HELENA CRISTINA MENDES BRÁS SILVA

MÉTODOS DE PARTIÇÃO E VALIDAÇÃO EM ANÁLISE CLASSIFICATÓRIA BASEADOS EM TEORIA DE GRAFOS

Departamento de Matemática Aplicada
Faculdade de Ciências da Universidade do Porto
Março / 2005

HELENA CRISTINA MENDES BRÁS SILVA

MÉTODOS DE PARTIÇÃO E VALIDAÇÃO EM ANÁLISE CLASSIFICATÓRIA BASEADOS EM TEORIA DE GRAFOS

Tese submetida à Faculdade de Ciências da Universidade do Porto
para obtenção do grau de Doutor em Matemática Aplicada

Sob a orientação de
Maria Paula de Pinho de Brito Duarte Silva
Sob a co-orientação de
Joaquim Fernando Pinto da Costa

À memória do meu
Querido Pai

Agradecimentos

Ao finalizar a elaboração desta dissertação gostaria de agradecer a todos os que contribuíram para que este trabalho fosse concluído.

À Professora Paula Brito, minha orientadora, expresso o meu sincero obrigado pelas orientações científicas, pelo apoio, pela dedicação e pela grande disponibilidade que sempre me dispensou.

Ao Professor Joaquim Pinto da Costa, meu co-orientador, agradeço as orientações científicas prontamente prestadas, sempre que solicitadas.

Ao ISEP e ao PRODEP, agradeço a dispensa de serviço docente concedida durante três anos, e o financiamento prestado no pagamento de proprinas.

Aos meus colegas do Departamento de Matemática do ISEP, mas em especial à Dr^a Helena Vieira da Silva, quero expressar o meu obrigado pela compreensão e apoio que me manisfestou; à Dr^a Alzira Faria, agradeço-lhe sinceramente as suas palavras de incentivo e carinho que sempre me dispensou.

À Professora Fernanda Sousa e, em especial, à Dr^a Sara Tavares, agradeço a colaboração prestada para a elaboração desta dissertação.

Aos Professores Maria Halkidi, Sergios Theodoridis, George Karypis e Lawrence Hubert, agradeço terem-me disponibilizado alguns conjuntos de dados e artigos.

À minha família, principalmente à minha mãe, expresso o meu sincero obrigado pelo seu apoio e o seu trabalho, quando tive que descuidar dos meus deveres de dona-de-casa e de mãe. Ao meu marido, Domingos, agradeço sinceramente o apoio e incentivo constantes que me deu ao longo destes anos de trabalho. Finalmente queria agradecer às minhas filhas Patrícia e Leonor pela compreensão da menor disponibilidade da mãe, principalmente nesta fase final.

Índice

Resumo

Aplicando a teoria de grafos à análise classificatória, é proposto um método de classificação não hierárquica validado por um índice de classificabilidade. O método pode ser aplicado a conjuntos de dados com atributos numéricos e categóricos, não requerendo qualquer pressuposto sobre a distribuição dos dados. O método de partição é baseado na coloração de grafos e usa um algoritmo sequencial estendido; o número e a composição das classes que melhor representam a estrutura existente no conjunto de dados, são aqueles que optimizam o valor do índice de classificabilidade. A ideia chave deste índice é a de que existem k classes compactas e bem separadas, se for possível definir um grafo *k-partite* completo nos dados, após a classificação. O índice de classificabilidade também permite identificar conjuntos de dados sem uma estrutura em classes.

É ainda proposto um índice de avaliação da qualidade de uma partição e alguns índices da avaliação da densidade e do isolamento de classes, que permitem seleccionar a melhor partição de entre as obtidas pelo método de classificação. Os atributos (ou variáveis) podem ser ainda considerados de modo diferenciado, obtendo-se partições em sub-espaços, permitindo que cada atributo ou bloco de atributos, tenha um contributo individual para a identificação de partições no conjunto de dados, tendo como base teórica os multigrafos. Neste contexto, são obtidas partições φ-projectadas resultantes de projecções dos multigrafos em grafos simples, ou partições de consenso obtidas a partir das partições associadas a cada um dos blocos de atributos. As distribuições teóricas dos índices apresentados são desenvolvidas usando modelos no contexto de grafos aleatórios.

A metodologia apresentada foi aplicada a conjuntos de dados reais e artificiais, comparando os resultados com os obtidos por outros métodos de classificação, constatando-se a eficiência do método, em particular, na detecção de classes de forma não esférica e sendo robusto no que diz respeito à presença de pontos isolados.

Abstract

Applying graph theory to clustering, we propose a partitional clustering method and a clustering tendency index, which may be applied to data sets with numerical and categorical values. No initial assumptions about the data set are requested by the method. The partitional algorithm is based on graph coloring and uses an extended greedy algorithm. The number of clusters and the partition that best fits the data set are selected according to the optimal clustering tendency index value. The key idea of this index is that there are k well-separated and compact clusters, if a complete k-partite graph can be defined in the data set, after clustering. The clustering tendency index can also identify data sets with no clustering structure.

We propose a partition quality evaluation index and some evaluation indices of clusters' density and isolation, which select the best partitions between the candidates resulting from the clustering method. The attributes (or variables) can be considered in a distinguishable way, obtaining partitions on sub-spaces, allowing each attribute or set of attributs, to have a individual role in the identification of clusters. This approach has the multigraphs as theoretical platform. In this context, we can obtain φ-projected partitions resulting from projections of multigraphs on simple graphs, or consensus partitions obtained from the partitions associated to each set of attributes.

We have applied our methodology on simulated and real data sets, so as to evaluate its performance, comparing it with other clustering methods. This study has showed that the method is efficient, in particular for detecting non spherical clusters and is robust as concerns outliers.

Résumé

Appliquant la théorie des graphes à la classification automatique, nous proposons une méthode de classification non hiérarchique, validée par un critère de classification. La méthode peut être appliquée à des ensembles de données décrites par des attributs numériques et catégoriques, et ne requiert pas d'hypothèse sur la distribution des données. La méthode de partitionnement est basée sur la coloration de graphes et utilise un algorithme séquentiel étendu; le nombre et la composition de classes déterminées sont ceux qui optimisent le critère de classification. Le principe du critère consiste à considérer qu'il existent k classes compactes et bien séparées, s'il est possible de définir un graphe *k-partite* complet sur les données, après la classification. Cette critère permet également d'identifier un ensemble de données sans structure en classes.

Nous proposons également un critère d'évaluation de la qualité de partitions et un couple de critères d'évaluation de la séparation et de la compacité des classes, qui permettent de choisir la meilleure partition parmi celles obtenues par la méthode de classification. Les attributs (ou variables) peuvent aussi être pris en compte séparément, permettant d'obtenir des classifications dans des sous-espaces, où chaque attribut ou bloc d'attributs, peut avoir un rôle individuel dans l'identification des classes. La plate-forme théorique de cette approche sont les multi-graphes. Dans ce contexte, nous pouvons obtenir des partitions φ-projetées résultant de la projection des multi-graphes en graphes simples, ou des partitions consensus obtenues à partir des partitions en sous-espaces associés à chaque bloc d'attributs. Les distributions théoriques des critères sont déterminées en utilisant des modèles de graphes aléatoires.

La méthodologie a été appliquée à des données réelles et artificielles, comparant les résultats avec les obtenus par d'autres méthodes de classification. On a pu constater l'efficacité de la méthode proposé, en particulier, dans la détection de classes de forme non sphérique; d'autre part, la méthode s'est montrée robuste en présence d'éléments isolés.

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

A análise classificatória é um ramo da análise de dados usado em muitos campos, tais como reconhecimento de padrões (aprendizagem não supervisionada), ciências biológicas e ecológicas (taxonomia numérica), ciências sociais (tipologia), teoria de grafos (*graph partitioning*), psicologia, etc. O principal objectivo da análise classificatória é formar subconjuntos, grupos ou estruturas, identificando classes que reflectem a organização de um conjunto de dados. As classes identificadas deverão obedecer a critérios de homogeneidade e de separação, traduzindo-se num maior grau de semelhança entre elementos da mesma classe, do que entre elementos de classes diferentes. A semelhança entre elementos depende da medida escolhida que condiciona a identificação das classes. Assim, o resultado de uma análise classificatória permite sumariar as relações entre um conjunto de elementos, representando-os por um menor número de classes de elementos [90].

O coração da análise classificatória é a selecção do método de classificação. O método a seleccionar deverá estar de acordo com o tipo de estrutura que se espera estar presente no conjunto de dados. Esta decisão é importante porque métodos de classificação diferentes, aplicados ao mesmo conjunto de dados, podem originar diferentes classificações. A maior parte dos métodos de classificação são de um dos seguintes tipos: *partição* ([154], [91] e [216]), *hierárquicos* ([135], [20], [91], [216] e [138]) e *fuzzy* ([55], [91], [216] e [166]). Novos métodos de classificação surgem frequentemente na literatura, reflectindo preocupações sobre aspectos não abordados na literatura clássica. Alguns aspectos focados são: um elevado número de elementos no conjunto de dados [228], a identificação de classes de formas arbitrárias [57], a presença de pontos isolados [57] a complexidade dos atributos que caracterizam os elementos do conjunto de dados

([100], [80] e [22]).

Os métodos de partição têm como objectivo decompor directamente o conjunto de dados num conjunto de classes disjuntas, obtendo-se uma partição que optimiza, pelo menos localmente, uma função de controlo. O algoritmo de partição mais conhecido é o *k-médias* de *MacQueen* [154] que pretende minimizar a soma dos quadrados das distâncias euclideanas de cada elemento ao centro da classe à qual pertence. Existem na literatura alguns trabalhos que têm como objectivo superar os inconveniente deste algoritmo nomeadamente, ter o número de classes como parâmetro de entrada ([190] e [185]), ser dependente da ordem de entrada dos dados [61] e não suportar atributos categóricos ([110] e [42]). Também tem sido superada a sua capacidade de lidar com um elevado número de elementos ([4],[185] e [125]). Em [188] é apresentada uma abordagem conceptual do algoritmo que incorpora descrições das classes obtidas de modo mais compreensível para os utilizadores.

Ainda no contexto dos algoritmos de partição, têm surgido na literatura alguns trabalhos nos quais os algoritmos são baseados em teoria de grafos ([226], [115], [105], [51] e [97]). De entre eles destaque-se o método de *Hansen & Delattre* [105] que é baseado na coloração de grafos e obtém classes de diâmetro (dissemelhança máxima entre pares de elementos da mesma classe) mínimo. Outros métodos pretendem em simultâneo a identificação de partições de diâmetro mínimo e de *split* (dissemelhança mínima entre elementos de classes diferentes) máximo [51]. Os algoritmos desenvolvidos no contexto de *graph partitioning* ([128] e [40]) pretendem dividir o conjunto dos vértices de um grafo num dado número de subconjuntos disjuntos, com um número de vértices em cada subconjunto aproximadamente igual; enquanto o número de arcos que unem vértices de conjuntos diferentes seja mínimo. Com o mesmo objectivo, os métodos espectrais ([53], [211], [124] e [176]) relacionam boas "partições" de um grafo aos valores e vectores próprios de certas matrizes derivadas do grafo.

Dada a diversidade de áreas onde se aplica a análise classificatória, existem algoritmos mais vocacionados para um determinado tipo de dados do que outros, não havendo, no entanto, nenhum algoritmo que seja melhor do que todos os outros quaisquer que sejam os elementos a analisar. Basicamente, a análise classificatória é um processo não supervisionado, não se prevendo a existência de qualquer conhecimento prévio da estrutura existente no conjunto de dados. Quase todos os algoritmos de classificação são fortemente dependentes das características do conjunto de dados, assim como dos

diferentes valores dos seus parâmetros de entrada. Assim, uma classificação obtida por algum algoritmo de classificação é localmente a melhor possível para os valores dos parâmetros de entrada atribuídos, mas pode não ser localmente a melhor classificação para o conjunto de dados em estudo. Outra questão particularmente preocupante é o facto de qualquer algoritmo de classificação produzir uma classificação no conjunto de dados, mesmo que este seja constituído por dados sem qualquer estrutura em classes. De facto, um qualquer critério de classificação impõe implicitamente um modelo nos dados de tal modo que quanto melhor se ajustarem a esse modelo, melhor é o desempenho do algoritmo que usa esse critério. Outro facto importante é que as classificações do mesmo conjunto de dados, obtidas usando critérios de classificação diferentes, podem diferir sobremaneira umas das outras. Pelo exposto, conclui-se que os algoritmos de classificação podem originar "falsas" classificações, sendo assim essencial que o resultado desses algoritmos seja sujeito a um processo de validação.

A validação em análise classificatória pode ser executada a três níveis. Primeiro deve-se verificar a existência de uma estrutura em classes no conjunto de dados. Se de facto existir tal estrutura, aplica-se um algoritmo de classificação, se não existir, a aplicação do algoritmo de classificação origina uma classificação que não reflecte a estrutura existente no conjunto de dados. O problema de verificar a existência ou não de uma estrutura em classes no conjunto de dados designa-se por *tendência de classificação* [216]. No procedimento de classificação segue-se a escolha de um "bom" algoritmo de classificação. Por exemplo, John W. Van Ness and Llyd Fisher ([65], [64] e [172]) desenvolveram um formalismo denominado *condições de admissibilidade*, e que consiste em onze critérios para avaliar algoritmos de classificação. A partir de alguma informação sobre os dados, o método indica quais os critérios de classificação que podem ser relevantes para a análise de um conjunto de dados em particular. Finalmente, deverá proceder-se à *avaliação quantitativa* do resultado do algoritmo de classificação escolhido. Assim, de acordo com o tipo de classificação usado, podemos validar partições (por exemplo [49], [113], [18], [183], [98] e [95]), partições *fuzzy* ([182], [223] e [222]), hierarquias ([216] e [91]) ou classes ([89], [168] e [91]).

Recorrendo a alguns conceitos da teoria de grafos, a presente dissertação tenta dar resposta a alguns problemas importantes em análise classificatória, destacando-se, pela sua importância, a determinação do número de classes do conjunto de dados, a identificação de conjuntos de dados sem uma estrutura em classes e a avaliação da qualidade dos resultados de uma classificação. Como já foi referido, estes problemas

têm tido cada vez mais atenção por parte dos investigadores, não se chegando, ainda, a resultados consensuais na comunidade científica. Este trabalho pretende contribuir para a procura de metodologias eficazes, no contexto dos problemas referidos. Grande parte dos trabalhos realizados com o objectivo de produzir algoritmos que sejam robustos perante pontos isolados e quanto à detecção de classes de formas não esféricas, dependem de alguns parâmetros de entrada. Pretende-se ainda que o método proposto nesta dissertação responda também a estes dois requisitos, não requerendo qualquer pressuposto sobre os dados.

A dissertação está organizada do seguinte modo.

O Capítulo 2 começa por referir alguns métodos de classificação clássicos, seguidos de alguns algoritmos mais recentes que pretendem colmatar algumas lacunas apresentadas pelos algoritmos clássicos. O capítulo segue com algumas generalidadas sobre validação em análise classificatória, concluindo com a referência a algumas medidas de comparação entre partições baseadas em pares de elementos concordantes.

O Capítulo 3 começa por referir alguns índices de validação existentes na literatura que têm como tarefa principal identificar o número de classes existente no conjunto de dados. O capítulo segue com a apresentação de um índice de validação denominado LCS ([141], [142] e [205]) e que associado ao algoritmo de classificação *k-prototype* [111], pretende identificar o número e a composição das classes existentes no conjunto de dados. O capítulo termina com a comparação do índice LCS com alguns índices referidos na secção anterior, quando aplicados a um conjunto de dados mistos.

O Capítulo 4 começa com uma breve descrição de alguns conceitos da teoria de grafos, necessários para uma boa compreensão do trabalho apresentado ao longo da dissertação. O capítulo segue com a descrição do método proposto, que consiste num algoritmo de partição baseado na coloração de grafos, validado por um índice de classificabilidade desenvolvido usando conceitos de grafos *k-partite* completos. A eficácia e a estabilidade do método são constatados em alguns conjuntos de dados artificiais. Finalmente, a complexidade de execução do método é avaliada, analisando o tempo de execução médio e o tempo de execução no pior caso.

Usando ainda conceitos de teoria de grafos, no Capítulo 5 é proposto um índice de

avaliação da qualidade de uma partição e alguns índices da avaliação da densidade e do isolamento de classes, que permitem seleccionar a melhor partição de entre as obtidas pelo método de classificação. Os índices são aplicados a conjuntos de dados artificiais de atributos (ou variáveis) numéricos e mistos, e comparados com alguns índices existentes na literatura. Um dos índices de isolamento apresentados é usado na definição de uma medida de dissemelhança entre classes, que permite definir grafos ponderados a usar numa abordagem hierárquica do método. São apresentados alguns exemplos que constatam a eficiência desta nova abordagem.

No Capítulo 6, os atributos (ou variáveis) que caracterizam os dados vão ser considerados de modo diferenciado, quer quanto ao tipo, quer quanto ao modo como cada atributo participa no processo de classificação. A diferenciação dos atributos tem como suporte teórico os multigrafos. Com esta abordagem obtém-se partições em sub-espaços, a partir das quais se obtêm partições φ-projectadas resultantes de projecções dos multigrafos em grafos simples, ou partições consenso obtidas a partir das partições associadas a cada um dos blocos de atributos. Estas novas abordagens serão testadas em alguns conjuntos de dados reais.

A avaliação do valor de uma medida de validação de uma estrutura em classes, exige uma comparação com o valor do índice obtido sob um modelo nulo de ausência de estrutura. O uso de grafos como base teórica de suporte ao desenvolvimento dos índices definidos nos capítulos anteriores, facilita o estudo estatístico dos mesmos, permitindo a definição das suas distribuições sob modelos definidos no contexto da teoria de grafos aleatórios. Assim, no Capítulo 7 são aplicados modelos da teoria de grafos aleatórios para definir as distribuições teóricas dos índices não normalizados e quando aplicados a partições não ajustadas.

No Capítulo 8, avalia-se o desempenho dos métodos propostos, quando aplicados a conjuntos de dados reais e artificiais, comparando-os com outros métodos de classificação.

Finalmente a dissertação é concluída, retomando as contribuições principais do presente trabalho e apontando vias para futuros desenvolvimentos.

Capítulo 2

Métodos e Validação em Análise Classificatória

A produção científica efectuada nos últimos anos no contexto da análise classificatória, reflecte uma grande preocupação em analisar conjuntos de dados cada vez mais complexos, quanto ao número de elementos a classificar, ao tipo de atributos (ou variáveis), ao formato das classes ou à presença de pontos isolados. A eficiência dos algoritmos, isto é, a capacidade de obter resultados com o menor dispêndio de recursos físicos e de tempo, é frequentemente provada para conjuntos de dados pesados, quer a nível do número de elementos, quer a nível do tipo de atributos. No entanto, nem sempre a eficácia dos algoritmos, ou seja, a segurança de se obter um bom resultado, é também alvo de estudo por parte dos investigadores. Verificando-se que qualquer algoritmo aplicado a qualquer conjunto de dados produz uma classificação, mesmo que os elementos não estejam organizados em classes, é de extrema importância que o resultado obtido seja alvo de algum tipo de validação.

2.1 Métodos de Classificação

A maior parte dos métodos de classificação são de um dos seguintes tipos: *partição* ([154], [91] e [216]), *hierárquicos* ([135], [20], [91], [216] e [138]) e *fuzzy* ([55], [91], [216] e [166]). Os métodos de partição têm como objectivo decompôr directamente o conjunto de dados num conjunto de classes disjuntas, obtendo-se uma partição que optimiza uma função de controlo. Os métodos hierárquicos criam uma decomposição hierárquica do conjunto de dados, construindo uma estrutura encaixada na qual as

classes ou são disjuntas, ou estão contidas umas nas outras. Os algoritmos *fuzzy* permitem que os elementos pertençam a mais do que uma classe, associando a cada atribuição um determinado peso.

Nesta secção serão referidos alguns algoritmos clássicos de cada um dos tipos de métodos de classificação, assim como outros emergentes na literatura mais recente, que tentam colmatar algumas lacunas apresentadas pelos algoritmos clássicos.

2.1.1 Métodos de Partição

Os métodos não hierárquicos ou de partição, têm como objectivo a obtenção de partições sobre o conjunto a classificar. De acordo com o modo de funcionamento, estes métodos distinguem-se entre *métodos centróides* e *métodos de transferências*.

2.1.1.1 Métodos de Partição Clássicos

Em geral, os métodos de partição centróides aplicados a um conjunto Ω de n elementos, partem de um subconjunto de k pontos, considerados os centros das classes ou pólos de agregação, que permitem iniciar o processo de afectação (sendo, em geral, k um parâmetro de entrada do algoritmo). O método é iterado, recalculando-se os centros em cada etapa; os elementos do conjunto a classificar vão sendo reafectados em função da sua dissemelhança aos centros. O critério de paragem é a não modificação da partição em duas etapas sucessivas.

Dois dos métodos centróides mais conhecidos são os métodos de *Forgy* [68] e o das *k-médias* de *MacQueen* [154]. O método de *Forgy* considera como centros das classes os centros de gravidade das classes que se mantêm constantes ao longo de cada iteração. O método consiste fundamentalmente nos seguintes passos:

1. Selecção de k elementos do conjunto Ω, sendo estes os centros iniciais das k classes;
2. Afectação dos restantes elementos às classes de cujos centros são mais semelhantes;
3. Determinação dos centros de gravidade das classes, sendo os novos centros das classes (já não são elementos de Ω);

4. Repetição dos passos (2) e (3) até que em duas iterações sucessivas se obtenha a mesma partição.

O método das *k-médias* de *MacQueen* considera como centro de cada classe o seu centro de gravidade, que é recalculado após cada afectação de um elemento a uma classe. A diferença fundamental entre estes dois métodos, consiste no facto de que no método de *MacQueen* os centros das classes são sistematicamente actualizados, enquanto que para o método de *Forgy* os centros mantêm-se constantes ao longo de cada iteração. Assim, o método de *MacQueen* consiste nos seguintes passos:

1. Selecção de k elementos do conjunto Ω, sendo estes os centros iniciais das k classes;
2. Para cada um dos $n - k$ restantes elementos:
   (a) Afectação do elemento à classe de cujo centro é mais semelhante;
   (b) Actualização do centro de gravidade da classe;
3. Após a classificação de todos os elementos, o método pode prosseguir seguindo um processo de trocas de elementos entre classes, recalculando, após cada transferência, o centro da classe origem e da classe destino.

Tal como para o método anterior, o critério de paragem é a não modificação da partição em duas etapas sucessivas.

A medida de dissemelhança utilizada para os dois métodos é, em geral, a distância euclideana.

Os métodos centróides pressupõem conhecido *a priori* o número de classes k, do conjunto de dados. No entanto, no caso do método das *k-médias*, *MacQueen* propôs uma variante do método em que há um aumento ou diminuição do número de classes k ao longo do processo, guiado pelas características dos dados. O processo depende da fixação de dois parâmetros: o parâmetro de refinamento R e o parâmetro de fusão C, que são usados para ajustar o número de classes. O processo de afectação de um elemento a uma classe é condicionado pelos parâmetros R e C do seguinte modo:

1. Processo de Refinamento: Se a dissemelhança de um elemento ao centro mais semelhante for superior ao parâmetro R, o elemento não é afectado e é tomado como centro de uma nova classe, aumentando-se deste modo o número de classes k;

2. Processo de Fusão: Para a partição obtida em cada etapa, determina-se o par de centros mais semelhantes, se a dissemelhança entre os centros for inferior ao parâmetro C, as respectivas classes são reunidas e o centro calculado. Este procedimento é repetido até que todos os centros distem entre si de pelo menos C. Assim, em cada etapa o número de classes k pode ser, eventualmente, diminuído.

Existem vários métodos para escolher os primeiros k centros ou representantes das classes.

A opção mais simples consiste em tomar para centros iniciais os k primeiros elementos do conjunto, sendo esta a escolha clássica para o método das *k-médias* de *MacQueen*. Outra opção consiste em numerar os indivíduos do conjunto de 1 a n e tomar os de ordem $\frac{n}{k}, \frac{2n}{k}, \ldots, \frac{(k-1)n}{k}$ e n. Este método é tão simples como o primeiro, tentando compensar a eventual tendência de agregação dos elementos segundo a ordem de observação, uma vez que esta ordenação pode influenciar de modo significativo o resultado final.

O algoritmo de *Forgy*, propõe que os elementos do conjunto sejam repartidos em k classes disjuntas, sendo os centros de gravidade destas classes os k pólos procurados. Trata-se, neste caso, de pontos artificiais (e não elementos do conjunto).

Uma opção mais eficiente mas mais dispendiosa é proposta em [136], e consiste numa prévia classificação ascendente hierárquica sobre um subconjunto do conjunto de dados. Os núcleos a tomar são os centros de gravidade das classes da partição em k classes fornecida pelo método hierárquico.

Em [180] é proposto um novo método para seleccionar os primeiros k centros: o método *wise-choice*. Este método considera como centros iniciais, elementos do conjunto de dados bem separados, centrados em vizinhanças densas. Para cada $x \in \Omega$, a *L-vizinhança* de x, é o conjunto $V(x)$ dos L elementos de Ω mais próximos de x, ou seja, são os elementos y que correspondem aos L maiores valores da função semelhança $s(x, y)$ para $y \in \Omega$. O primeiro centro c_1 é o elemento a que corresponde a vizinhança mais densa, ou seja, é o elemento x que obedece à condição:

$$max_{\{x \in \Omega\}} \sum_{y \in V(x)} s(x, y) \tag{2.1}$$

Cada um dos restantes centros, $c_i, i = 2, \ldots, k$, será o elemento x, de entre os ainda não seleccionados como centros, que maximiza a condição:

$$min_{\{1 \leq j \leq i-1\}} \frac{\sum_{y \in V(x)} s(x,y)}{s(x,c_j)} \tag{2.2}$$

onde C é o conjunto de centros já seleccionados.

Finalmente, se existir alguma informação *a priori* sobre as classes, pode optar-se por ser o utilizador a escolher os centros iniciais.

O representante de cada classe poderá ser um elemento artificial dado pelo centro de gravidade (algoritmos *Forgy* e *k-médias* de *MacQueen*), por um elemento central do conjunto de dados, usando o elemento mais central da classe (*Partitioning Around Medoids, PAM* [131] e [58]), ou o vector de modas dos atributos [42]. No entanto, os representantes das classes podem ser mais complexos. Por exemplo, no *método das nuvens dinâmicas* [52], é definida uma representação simbólica de cada classe, chamada *núcleo* e que poderá ser um ponto, um conjunto de pontos, uma função, etc. Este método consiste em, a partir de uma configuração inicial de núcleos, construir uma partição por afectação dos elementos, e, a partir da partição obtida, estabelecer uma nova representação por um conjunto de núcleos, que permitirão, por sua vez, gerar uma nova partição. Assim, o processo consiste na aplicação de uma etapa de representação seguida de uma etapa de afectação de modo iterativo até à convergência. Sob certas condições nas funções que permitem afectar os elementos às classes e calcular os núcleos, este método faz decrescer um critério W que mede a adequação entre as classes e os respectivos núcleos. Seja $\mathbb{C}_k$ o conjunto das partições de Ω em k classes, $\mathbb{C} = \{C_1, ..., C_k\}$, e $\mathbb{N}_k$ o conjuntos dos k-uplos $\mathbb{N}$ de núcleos admissíveis. O critério a minimizar consiste em: $W(\mathbb{N}, \mathbb{C}) = \sum_{i=1}^{k} D(N_i, C_i)$, onde D é uma medida de dissemelhança que avalia o ajuste do núcleo N_i à classe C_i.

Os métodos de transferências partem de uma partição inicial prosseguindo de modo iterativo, transferindo elementos entre classes sempre que o critério adoptado de melhoria de partição é verificado. O processo termina quando nenhuma transferência conduz a uma melhoria do critério. A partição obtida corresponde, no caso geral, apenas a um óptimo local. O critério a melhorar em cada iteração pode ser, por exemplo, a soma dos resíduos, isto é, a soma dos quadrados dos desvios dos elementos

aos centros das respectivas classes (*método de Beale* [15] e [179]). Pode ainda optar-se por um critério mais sofisticado resultado da soma das compacidades das classes de uma partição central relativamente a uma família de partições induzidas pelas variáveis descritoras (*método das partições centrais de* Régnier [191]).

Em geral, os métodos de partição convergem para um óptimo local, atingido quando em duas iterações sucessivas se obtém a mesma partição. No caso do método das núvens dinâmicas, variante dos centros de gravidade (método de *Forgy*), em cada iteração, o valor da inércia intraclasse diminui, tendo então o método tendência a procurar classes circulares, de volumes iguais e de baixa inércia.

Refira-se ainda um método de partição clássico, que aplicado a conjuntos de dados com uma única variável quantitativa ou qualitativa ordinal, encontra uma partição óptima global em k classes: O algoritmo de *Fisher* [66]. O algoritmo consiste em calcular por recorrência uma sucessão de partições óptimas, das quais é extraída a partição óptima em k classes.

Em geral, os algoritmos de partição são de ordem linear, possibilitando a sua aplicação a conjuntos de dados com elevado número de elementos. No entanto, dependem da ordem de entrada dos dados, são em geral incapazes de lidar com pontos isolados e de identificar classes de forma não convexa e finalmente, requerem alguns pressupostos iniciais, usualmente o número de classes que se julgue existir no conjunto de dados.

Face aos inconvenientes inerentes a este tipo de algoritmos, têm surgido na literatura alguns métodos que pretendem minimizar alguns dos problemas apresentados pelos algoritmos de partição clássicos.

2.1.1.2 Métodos de Partição baseados no Método das *k-médias*

O algoritmo de partição mais conhecido é o *k-médias* de *MacQueen* [154] que pretende minimizar a soma dos quadrados das distâncias euclideanas de cada elemento ao centro da classe à qual pertence. Existem na literatura alguns trabalhos que têm como objectivo superar os inconvenientes deste algoritmo nomeadamente, ter o número de classes como parâmetro de entrada ([190] e [185]), ser dependente da ordem de entrada dos dados [61] e não suportar atributos categóricos ([110] e [42]). Também tem sido

superada a sua capacidade de lidar com um elevado número de elementos ([4],[185] e [125]). Em [188] é apresentada uma abordagem conceptual do algoritmo que incorpora descrições das classes obtidas de modo mais compreensíveis para os utilizadores.

O algoritmo *evidence accumulation clustering* ([75] e [76]) é um algoritmo de partição e pretende que a sua execução seja independente dos pressupostos iniciais, nomeadamente do número de classes, e que a sua eficácia não se limite na identificação classes de forma esférica. O algoritmo consiste na procura de classes consistentes de entre um conjunto de partições do conjunto de dados. A abordagem utilizada engloba a atribuição de associações entre pares de elementos, dadas pelo número de vezes que eles ocorrem na mesma classe nas diversas partições do conjunto de dados, obtidas por execuções independentes do algoritmo das *k-médias* para diferentes inicializações aleatórias. A matriz de ocorrências construída pelas associações entre pares de elementos representa uma nova medida de semelhança, e serve de suporte para a identificação de classes consistentes usando um algoritmo baseado numa árvore de comprimento mínimo (*minimum spanning tree*).

O algoritmo *Clustering Large Applications based on Randomized Search* (CLARANS) [177] é do tipo *k-medoids* e tem como objectivo lidar de modo mais eficiente com pontos isolados, ser independente da ordem de entrada dos dados e analisar conjuntos com um elevado número de elementos. Inicialmente, este algoritmo escolhe de modo arbitrário um conjunto de k elementos, como representantes das k classes pretendidas. Após a atribuição de todos os elementos às classes de cujos representantes (*medoids*) são mais semelhantes, os representantes de cada classe são substituídos por outros elementos da classe, de entre um subconjunto de elementos escolhidos aleatoriamente, se essa substituição resultar numa melhoria da qualidade da classificação. Esta medida da qualidade da partição é baseada na dissemelhança média entre cada elemento e o *medoid* da classe à qual pertence.

Em [165] é apresentada uma nova versão do algoritmo das *k-médias*, na qual são associados pesos aos atributos, dependentes da classe à qual o elemento pertence, de modo a reduzir ou transformar o espaço de atributos.

2.1.1.3 Métodos de Partição baseados em Densidade

Os algoritmos deste tipo formam classes, procurando vizinhanças de cada elemento, que obedeçam a certos requisitos iniciais. Em geral, estes algoritmos detectam classes de forma arbitrária e são eficientes em conjuntos de dados com um elevado número de elementos.

O algoritmo de classificação *Density Based Spatial Clustering of Applications with Noise* (DBSCAN) [57] pretende identificar classes de forma arbitrária, ser robusto relativamente a pontos isolados e apresentar um bom desempenho em conjuntos de dados com um elevado número de elementos. A estratégia seguida consiste em detectar zonas densas (classes), agrupando elementos vizinhos na mesma classe de acordo com critérios específicos, os quais exigem que a uma vizinhança de cada elemento com um determinado raio, deverão pertencer um número mínimo de elementos da mesma classe. A execução do algoritmo é condicionada por alguns parâmetros de entrada. O algoritmo IncrementalDBSCAN [56] usa o DBSCAN como base para efectuar a classificação, fazendo as alterações convenientes para uma aplicação incremental, tendo em conta a sua aplicação a conjuntos de grandes quantidades de dados actualizados frequentemente. A abordagem incremental permite que as actualizações sejam feitas somente nas classes onde os elementos são inseridos ou de onde são retirados.

O algoritmo *Distributed Based Clustering of Large Spatial Databases* (DBCLASD) [224] determina de modo dinâmico o número e a forma das classes, sem requerer qualquer parâmetro de entrada e é eficiente para conjuntos de dados com um elevado número de elementos. Este algoritmo baseia-se no pressuposto de que os elementos no interior de uma classe são uniformemente distribuídos, e realiza uma análise da distribuição de probabilidade da função distância ao vizinho mais próximo dos elementos de cada classe. O algoritmo DBCLASD é um algoritmo incremental, no sentido em que a atribuição de um elemento a uma classe é baseada nos elementos processados até ao momento, sem considerar todos os elementos do conjunto de dados ou mesmo todos os elementos da classe respectiva. O algoritmo aumenta de modo incremental uma classe inicial com os seus elementos vizinhos, até que o conjunto de distâncias dos elementos aos vizinhos mais próximos da classe resultante, deixe de obedecer à função distribuição esperada.

2.1.1.4 Métodos de Partição para Atributos Mistos

Todos os algoritmos descritos até ao momento lidam só com atributos numéricos. A complexidade dos atributos de conjuntos de dados mais sofisticados, requerem algoritmos de classificação que suportem outro tipo de atributos, nomeadamente aqueles que não obedecem a uma ordenação numérica. Numa primeira abordagem refira-se o trabalho descrito em [188] no qual os atributos categóricos são convertidos em atributos binários (usando 1 para representar uma categoria presente e 0 para representar uma categoria ausente), e tratados como atributos numéricos no algoritmo das *k-médias*. Esta abordagem falha uma vez que qualquer operação numérica com os valores binários perde o significado, nomeadamente no cálculo dos representantes de cada classe. No entanto, a transformação de atributos categóricos em atributos numéricos pode ser eficiente em alguns casos. Por exemplo, no trabalho referido em [67] no contexto de processamento de imagem e percepção de cor, cada cor é representada por um terno de valores reais com as intensidade RGB ("Red", "Green" e "Blue"). Qualquer operação numérica realizada com estes atributos categóricos, resulta num valor válido, ou seja numa cor.

Constatando que a representação numérica de atributos categóricos pode mascarar a estrutura existente no conjunto de dados, têm surgido na literatura alguns trabalhos que desenvolvem funções de dissemelhança ou semelhança adequadas a atributos não numéricos. Por exemplo, no contexto dos algoritmos de partição surgiu o algoritmo *k-mode* [110] que realiza uma extensão do algoritmo das *k-médias* de suporte a atributos categóricos. Este algoritmo introduz uma nova medida de dissemelhança para elementos com atributos categóricos, substituindo as médias das classes por modas, e usando um método baseado em frequências para actualizar as modas no processo de classificação. O algoritmo *k-prototype* [111] é uma extensão do algoritmo *k-mode* suportando atributos mistos (numéricos e/ou categóricos). Neste algoritmo, a dissemelhança entre dois elementos é dada pela soma do quadrado da distância euclideana entre os atributos numéricos, e o número de categorias não coincidentes para cada um dos atributos categóricos. As duas componentes da medida de dissemelhança são balançadas por um parâmetro que evita o favorecimento de um dos dois tipos de atributos.

Em [118] é apresentada uma generalização das métricas *Minkowski* a aplicar a elementos caracterizados por atributos mistos, sendo a sua eficácia constatada no contexto da classificação hierárquica conceptual e da análise em componentes principais. Em [221] são apresentadas algumas funções de dissemelhança que são diferenciadas de acordo

com o tipo de atributos a medir, podendo ser numéricos e/ou categóricos. Apesar de as funções terem sido desenvolvidas no contexto da análise discriminante, também podem ter aplicabilidade em análise classificatória, uma vez que medem a dissemelhança entre dois elementos do conjunto de dados com o objectivo, ou de formar classes, ou de classificar um elemento em uma das classes existentes. Uma abordagem baseada em sistemas dinâmicos para classificar conjuntos de dados com atributos categóricos é apresentada em [84], no qual foi definida uma função de semelhança baseada nas ocorrências de categorias dos diferentes atributos categóricos, aos quais são atribuídos pesos. O trabalho proposto consiste numa generalização da metodologia *spectral graph partitioning* ([53] e [62]), que consiste numa generalização da noção de vectores próprios não principais, para sistemas dinâmicos não lineares.

2.1.1.5 Métodos de Partição baseados em Teoria de Grafos

Existem na literatura alguns algoritmos de partição baseados em teoria de grafos, que são aplicados ao grafo representativo do conjunto de dados, fazendo-se uma correspondência directa entre os vértices do grafo e os elementos a classificar (ver [226] e [115] para uma revisão).

Considere-se o diâmetro de uma partição a dissemelhança máxima entre pares de elementos da mesma classe, e *split* de uma partição a dissemelhança mínima entre elementos de classes diferentes. Alguns algoritmos pretendem encontrar partições de diâmetro mínimo [105], enquanto outros pretendem em simultâneo a identificação de partições de diâmetro mínimo e de *split* máximo [51]. No primeiro caso, o algoritmo é baseado na coloração de grafos e prova-se que dado o número de classes k, as classes de vértices coloridos com a mesma cor numa coloração óptima em k cores, correspondem a uma partição em k classes de diâmetro mínimo. No segundo caso, no método de classificação de *Delattre & Hansen* são usados critérios orientadores da identificação das classes que procuram optimizar os dois aspectos considerados. O primeiro critério consiste na razão *diâmetro/split* que quanto menor for, melhor é a partição; correspondendo a uma partição óptima para o caso em que essa razão é menor ou igual a 1. A *qualidade de homogeneidade* (q_h) de uma partição é baseada no número de pares de elementos pertencentes a classes diferentes, com dissemelhança maior ou igual ao diâmetro da partição. A *qualidade de isolamento* (q_s) de uma partição é baseada no número de pares de elementos pertencentes à mesma classe, com dissemelhança menor ou igual ao *split* da partição correspondente. A execução

do algoritmo é orientada com o objectivo de maximizar q_h e maximizar q_s.

Em [97] é apresentado um algoritmo de classificação não hierárquica baseado numa função densidade, aplicada a cada um dos vértices do grafo parcial, representativo do conjunto de dados, definido num espaço métrico (qualitativo, quantitativo ou binário). O método pretende extrair classes parciais (nas quais nem todos os elementos são classificados) ou estabelecer partições (onde todos os elementos são classificados em classes disjuntas), procurando partes conexas que apresentem valores altos para a função de densidade. O grafo parcial é construído de forma a que dois vértices sejam adjacentes se a dissemelhança entre eles for menor que um dado valor limite. Para cada vértice do grafo, a função densidade baseia-se na sua vizinhança e pretende-se que apresente valores altos quando os elementos vizinhos do vértice estiverem próximos (baixa dissemelhança). Foram apresentadas três funções de densidade baseadas na média, no mínimo e no máximo das distâncias entre cada vértice e todos os seus vizinhos.

Os algoritmos desenvolvidos no contexto de *Graph partitioning* ([128] e [40]) pretendem agrupar o conjunto dos vértices de um grafo num dado número k de subconjuntos disjuntos, de modo a minimizar o número de arcos entre subconjuntos. Exemplos desta metodologia são o algoritmo AutoPart [40] e o algoritmo METIS ([129], [130] e [128]). O método AutoPart não requer qualquer parâmetro de entrada, e propõe agrupar vértices de um grafo em classes. O método segue a ideia de que é possível realizar uma reordenação da matriz adjacente de modo a que vértices "similares" sejam agrupados conjuntamente. Assim, a matriz adjacente pode consistir em blocos rectangulares ou quadrados homogéneos de alta (baixa) densidade, representando o facto de que alguns grupos de vértices têm mais (menos) ligações a outros grupos. Quanto ao algoritmo METIS tem como objectivo agrupar os vértices de um grafo em subconjuntos proximadamente de igual tamanho, minimizando o número de arcos entre vértices de subconjuntos diferentes.

Ainda no contexto da metodologia *Graph partitioning* refiram-se os métodos espectrais ([53], [211], [124] e [176]) que relacionam boas "partições" de um grafo aos valores e vectores próprios de certas matrizes derivadas do grafo. Sejam $m_1 \geq m_2 \geq \ldots \geq m_k$ o número de elementos dos k subconjuntos e $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k$ os k maiores valores próprios de uma matriz que resulta da soma da matriz adjacente do grafo com uma matriz diagonal, tal que a soma de todos os elementos da matriz soma seja zero. Seja

E_c o número de arcos que unem vértices de conjuntos diferentes, pretende-se que E_c seja mínimo. Em [53] prova-se que $|E_c| \geq -\frac{1}{2}\sum_{r=1}^{k} m_r \lambda_r$.

2.1.2 Métodos Hierárquicos

Os algoritmos do tipo hierárquico criam uma decomposição hierárquica do conjunto de dados, procedendo sucessivamente ou reunindo classes mais pequenas em classes maiores (seguindo uma estratégia aglomerativa ou ascendente), ou dividindo classes maiores em classes mais pequenas (seguindo uma estratégia descendente). O resultado do algoritmo hierárquico é uma árvore de classes encaixadas, na qual cada nó da árvore representa uma classe e as classes ou são disjuntas ou estão contidas umas nas outras. Cortando a árvore num determinado nível, obtém-se uma partição do conjunto de dados. Basicamente, os diferentes métodos hierárquicos diferem entre si pelo método usado para o cálculo da dissemelhança entre classes, que condiciona a sua reunião ou divisão. Exemplos típicos de algoritmos hierárquicos ascendentes [135] são os baseados nos critérios do índice do mínimo, do índice do máximo, do índice da média, e do índice de Ward (em [20] é apresentado um trabalho de comparação entre estes quatro algoritmos) e o algoritmo CHAVL ([143] e [144]). Este tipo de algoritmos, sendo eficientes na detecção de classes de formas arbitrárias (por exemplo, algoritmo hierárquico ascendente com o critério do índice do mínimo) e sendo robustos quanto à presença de pontos isolados, têm como principal inconveniente o tempo de execução não linear e o enorme custo de entradas/saídas para grandes conjuntos de dados. O desempenho deste tipo de algoritmos é condicionado pelo facto de que em cada etapa, o critério associado ao algoritmo é calculado para todos os pares de elementos do conjunto e, nalguns casos, a matriz de dissemelhança entre classes deverá estar disponível em memória. O algoritmo baseado em *vizinhos recíprocos* ([187] e [17]) tenta melhorar o desempenho dos algoritmos hierárquicos ascendentes, limitando o número de comparações em cada etapa do processo. Recentemente têm surgido na literatura outros trabalhos com o objectivo de superar este inconveniente, adoptando algumas estratégias computacionais para optimizar o desempenho deste tipo de algoritmos. Alguns exemplos destes algoritmos mais recentes são o *BIRCH* ([229] e [228]), o CURE [99] e o CLIQUE [2]. Alguns algoritmos tais como o ROCK [100] e o CACTUS [80] são algoritmos hierárquicos que suportam dados categóricos.

2.1.2.1 Métodos Hierárquicos para Conjuntos de Dados com um Elevado Número de Elementos

O algoritmo *Balanced Iteractive Reducing and Clustering using Hierarchies* (BIRCH) ([229] e [228]) é um método de classificação especialmente desenvolvido para conjuntos de dados com um elevado número de elementos definidos num espaço métrico. Tendo em conta a limitação dos recursos físicos, nomeadamente de memória, este algoritmo foi especialmente desenvolvido para os casos em que o espaço em memória é insuficiente para conter o conjunto de dados. Inerente a este tipo de solução estão os custos de entradas/saídas que se pretendem mínimos. O algoritmo começa por gerar em memória um sumário compacto do conjunto de dados, pretendendo reter a maior quantidade possível de informação sobre a distribuição dos elementos, identificando à partida zonas densas e zonas esparsas. O *Clustering Feature* (CF) é o conceito base do algoritmo BIRCH e consiste num terno associado a cada classe, que é armazenado em memória numa árvore balanceada de actualização incremental, e contém toda a informação necessária para o cálculo de todas as medidas de suporte às decisões de classificação, requeridas pelo algoritmo. Com esta abordagem é feita uma grande economia do espaço em memória, tendo em conta que não são os próprios elementos pertencentes a cada uma das classes que são armazenados, mas sim os sumários que contêm o número, a soma e a soma dos quadrados dos valores dos atributos em cada uma das classes. Após a pré-classificação usa-se um qualquer método de análise classificatória, aplicado aos sumários carregados em memória e não ao conjunto de dados completo, atribuindo importância diferente aos elementos do conjunto de dados, sendo aqueles que se encontram juntos em zonas densas, tratados colectivamente e não individualmente. O algoritmo BIRCH trabalha com a memória disponível e a complexidade de entradas/saídas é um pouco maior que uma única passagem nos dados, sendo o tempo gasto em operações de entrada/saída linearmente proporcional ao tamanho do conjunto. Para a eficiência do BIRCH é necessário o estabelecimento de valores adequados para os parâmetros de entrada.

O algoritmo de classificação hierárquica *Clustering Using Representaives* (CURE) [99] aplica-se a conjuntos de dados definidos num espaço métrico, pretende detectar classes de qualquer forma com grande variação quanto ao número de elementos, sendo ainda robusto na presença de pontos isolados. Cada classe é representada por um número fixo de elementos, que são gerados seleccionando pontos "bem espalhados" pela classe e depois "estreitando-os" em direcção ao centro da classe, através de uma fracção específica. Estes pontos espalhados capturam a forma e a extensão das classes. Ter

mais que um ponto representativo de cada classe, permite que o algoritmo CURE se ajuste bem a formatos não esféricos e o estreitamento ajuda a atenuar o efeito de pontos isolados. Para lidar com conjuntos de dados com um elevado número de elementos, o algoritmo CURE selecciona uma amostra aleatória do conjunto de dados, tendo as devidas precauções para que a amostra tenha informação suficiente quanto à geometria das classes. Sobre esta amostra é realizada uma pré-classificação seguindo-se um processo hierárquico ascendente, pelo qual as classes cujos pares de representantes estejam mais próximos, são sucessivamente reunidas. Após a remoção dos pontos isolados é feita uma classificação final, pela qual os elementos não seleccionados na amostra, são afectadas às classes que lhe estão mais próximas.

2.1.2.2 Métodos Hierárquicos para Conjuntos de Dados com Elevada Dimensão

Na literatura têm surgido alguns algoritmos que tentam lidar eficientemente com conjuntos de dados com um elevado número de atributos. Por exemplo, o algoritmo *CLustering In QUEst* (CLIQUE) [2] usa uma aproximação baseada em densidade para a classificação, considerando uma classe como uma região com uma maior densidade de elementos que a região envolvente. O algoritmo pretende identificar classes densas em subespaços de dimensão máxima, fazendo projecções do conjunto de dados de entrada num subconjunto de atributos, tendo a propriedade de que estas projecções incluem regiões de alta densidade.

O algoritmo *Categorical Clustering Using Summaries* (CACTUS) [80] suporta atributos categóricos, encontra classes em subespaços de todos os atributos, gerando "candidatas" a classes encontradas em projecções de dimensão crescente dos atributos. Tal como o algoritmo BIRCH, este algoritmo considera que um sumário de todo o conjunto de dados é suficiente para determinar o conjunto de classes, possibilitando assim a sua aplicação a conjuntos de dados com um elevado número de elementos.

Em [155] a questão relativa à selecção de atributos é abordada no contexto da programação matemática, pela qual o problema é formulado com uma função objectivo paramétrica.

2.1.2.3 Um Método Hierárquico para Conjuntos de Dados com Atributos Mistos

O algoritmo *Robust Clustering using Links* (ROCK) [100] pertence à classe dos algoritmos hierárquicos ascendentes e trata elementos com atributos categóricos, apresentando uma nova abordagem em classificação baseada no conceito de *ligações* entre elementos numa vizinhança. O número de ligações entre um par de elementos é o número de vizinhos comuns, sendo vizinhos aqueles cuja função semelhança é maior ou igual a um dado parâmetro de entrada do algoritmo. Os elementos pertencentes a uma mesma classe têm, em geral, um elevado número de vizinhos comuns, e consequentemente mais ligações. Sendo assim, durante o processo de análise classificatória, se se reunirem em primeiro lugar classes/elementos com o maior número de ligações, obtemos um resultado melhor e classes com maior significado. Ao contrário de dissemelhanças ou semelhanças entre um par de elementos com propriedades locais inerentes, o conceito de ligação incorpora informação global àcerca dos outros elementos na vizinhança, tornando-se assim um método mais robusto. Quanto maior for o número de ligações entre dois elementos, maior é a probabilidade de eles virem a pertencer a uma mesma classe. Especificamente, este método começa por encontrar uma amostra aleatória do conjunto de dados, com os devidos cuidados quanto ao seu tamanho para que a qualidade da classificação não seja sacrificada. A este conjunto é aplicado um algoritmo de classificação hierárquica baseado em ligações, reunindo-se em cada passo as duas classes com o maior número de ligações entre elas. Finalmente, as classes que contêm elementos da amostra são aumentadas com os restantes elementos que não participaram na construção da hierarquia.

2.1.2.4 Um Método Hierárquico baseado em Densidade

No contexto da análise de imagem em [201] é apresentado um algoritmo hierárquico (GRIDCLUS) que usa uma grelha multidimensional para organizar o espaço à volta dos dados, agrupando-os em blocos rectangulares. A classificação é feita por um algoritmo baseado na procura da vizinhança topológica de cada elemento, usando a densidade de cada bloco, calculada pelo número de elementos nele contido e pelo seu volume. Os blocos com maior densidade tornam-se os centros das classes. Os restantes blocos são classificados interativamente de acordo com o valor de sua densidade, construindo novos centros de classes ou juntando-se a classes já existentes.

2.1.2.5 Métodos Hierárquicos baseados em Teoria Grafos

O algoritmo *EjCluster* [81] é um algoritmo hierárquico que surgiu no contexto do processamento de sinal, usa conceitos de teoria de grafos e gera automaticamente uma condição de paragem. Este algoritmo pretende determinar automaticamente o número e a composição das classes, detectando classes separadas não esféricas. É definido um grafo representativo dos dados no qual um caminho entre dois vértices v_i e v_j quaisquer, de comprimento ℓ, é dado por $\{v_i, v_{i+1}, \ldots, v_{i+\ell}\}$, tais que $v_{i+\ell} = v_j$; v_m para $m = 1, \ldots, n$ são os vértices do grafo correspondentes aos n elementos do conjunto a classificar e $\{v_m, v_{m+1}\}$ para $m = i, \ldots, i + \ell - 1$ são vértices adjacentes. O processo de classificação é baseado numa medida de dissemelhança entre elementos, dada pelo maior arco entre dois vértices consecutivos que pertencem ao caminho que une os dois elementos. A ideia principal é que dois elementos pertencem a uma mesma classe se for possível ir do primeiro para o segundo por passos "suficientemente pequenos". Este algoritmo é do tipo descendente, não necessita de nenhum parâmetro de entrada, tem um custo computacional superior ao algoritmo das *k-médias* e falha na detecção de classes não totalmente separadas.

Em [216] é apresentado um esquema genérico de execução de um algoritmo hierárquico ascendente baseado em grafos, nos quais as classes são identificadas como sendo subgrafos conexos. A identificação de classes válidas pode ainda estar condicionada por alguma propriedade $h(k)$, à qual os elementos do subgrafo deverão obedecer. Algumas propriedades típicas que poderão ser adoptadas são: (1) a *conectividade dos nós* de um subgrafo conexo é o maior inteiro k, tal que qualquer par de vértices estão ligados por pelo menos k caminhos sem vértices em comum; (2) a *conectividade dos arcos* de um subgrafo conexo é o maior inteiro k, tal que qualquer par de vértices estão ligados por pelo menos k caminhos sem arcos em comum e (3) o *grau de um subgrafo conexo* é o maior inteiro k, tal que cada vértice tem pelo menos k arcos incidentes. No processo aglomerativo, as classes (subgrafos conexos) são reunidas com base nas medidas de dissemelhança ou semelhança dos seus vértices, e condicionado pelo facto de que a classe ou subgrafo conexo resultante da união, ou tem a propriedade $h(k)$ ou é completo. Como exemplos deste tipo de abordagem considerem-se os algoritmos hierárquicos ascendentes baseados no índice do mínimo e no índice do máximo. O primeiro não requer qualquer propriedade a satisfazer, só exigindo que os subgrafos sejam conexos (condição mínima); quanto ao segundo também não requer qualquer propriedade, no entanto, exige que os subgrafos conexos sejam completos (condição máxima). Ou seja, o critério adoptado para a formação de uma nova classe é o mais

fraco possível no caso do índice do mínimo, sendo o mais forte possível para o índice do máximo. Existe uma variedade de algoritmos entre estes dois extremos, condicionados por diferentes escolhas da propriedade $h(k)$.

2.1.3 Metodologia VL - Validade da Ligação

A metodologia VL - Validade da Ligação - engloba um conjunto de métodos que recorrem a noções probabilísticas para definir as funções de comparação entre elementos e entre classes de um conjunto de dados.

Seja Ω o conjunto de n elementos a classificar, C_i e C_j duas classes com n_i e n_j elementos respectivamente, e c uma medida de comparação entre elementos. Nesta abordagem, os $\binom{n}{2}$ valores da medida de comparação entre dois quaiquer elementos do conjunto de dados, são consideradas realizações de uma variável aleatória Z. Seja γ uma medida de probabilidade para avaliar a ligação entre os elementos $x, y \in \Omega$, γ é dada por:

$$\gamma(x, y) = P(Z \leq c(x, y)) = F_Z(c(x, y)) \tag{2.3}$$

se c for uma medida de semelhança e por

$$\gamma(x, y) = P(Z \geq c(x, y)) = 1 - F_Z(c(x, y)) \tag{2.4}$$

se c for uma medida de dissemelhança. Considerando como hipótese de referência, independência entre os elementos de Ω, admite-se que o conjunto $\{\gamma(x, y) | x \neq y, x, y \in \Omega\}$ constitui uma amostra de $\binom{n}{2}$ variáveis aleatórias de lei Uniforme no intervalo $[0, 1]$. Assim, os valores $\gamma(x, y)$ são obtidos a partir da função de distribuição da função de comparação entre elementos. A partir da função de comparação entre elementos, foram criadas várias funções de comparação entre classes, a usar, quer num contexto hierárquico, quer num contexto não hierárquico.

Em [139], [143] e [144], *Lerman* propôs o método de classificação hierárquica *Agregação pela Verosimilhança da Ligação* - método AVL - que utiliza a função distribuição do máximo dos valores de comparação entre elementos, para definir a função de comparação entre classes, dada por:

$$\Gamma_{AVL}(C_i, C_j) = (\gamma_{Max}(C_i, C_j))^{n_i n_j} \tag{2.5}$$

Foi posteriormente definida [7] uma família paramétrica de métodos associados ao método AVL, em que a função de comparação entre classes é dada por:

$$\Gamma(C_i, C_j) = (\gamma_{Max}(C_i, C_j))^{1+\varepsilon[(n_i n_j)^{\xi}-1]} \tag{2.6}$$

destacando-se o critério AVB, que se obtém para $\varepsilon = 1$ e $\xi = 0.5$.

O desempenho deste método foi estudado por Bacelar-Nicolau [6]. Em [178] Nicolau propôs um novo método - método AVM - que utiliza a função distribuição da média aritmética dos valores de comparação entre elementos. Mais recentemente, foram propostos os métodos de classificação ascendente hierárquica baseados nas estatísticas da mediana, do centro e da média geométrica [210].

No contexto da classificação não hierárquica, Nicolau e Brito ([180] e [181]) consideraram como critério de afectação dos elementos às classes, a função distribuição da média aritmética dos valores de comparação entre elementos a afectar e os membros da classe. Este método denominado NHMEAN surge como uma adaptação do método AVM de classificação hierárquica [178], e foi aprofundado em [36] e [181] mostrando que, ao contrário do método AVL, fornece bons resultados quando aplicado ao quadro não hierárquico. Mais recentemente, em [37] foram propostos novos métodos, que consideram como critério de afectação dos elementos às classes, as funções de distribuição da mediana, da média geométrica e do centro, dos valores de comparação entre elementos a afectar e os membros da classe - métodos NHMED, NHVGM e NHVC, respectivamente.

Em [210] são apresentadas algumas experiências de comparação dos diferentes critérios da metodologia VL em classificação hierárquica, das quais se conclui que a capacidade destes métodos de recuperar a estrutura dos dados depende fortemente do tipo e da intensidade da estrutura dos dados. No entanto, e em geral, o método baseado na estatística da mediana mostrou um bom desempenho na recuperação da estrutura subjacente aos dados.

2.1.4 Partições *Fuzzy*

Considere-se uma partição do conjunto $\Omega = \{x_1, \ldots, x_n\}$ em k classes, $C_1, \ldots, C_k$. As classes podem ser representadas pelas funções de *pertença* ([222] e [216]), $u_i : \Omega \rightarrow \{0, 1\}$ tais que $u_i(x) = 1$ se $x \in C_i$ e $u_i(x) = 0$ se $x \notin C_i$, para $i = 1, \ldots, k$. A representação de uma partição *fuzzy* de Ω em k classes, pode obter-se generalizando as funções u_i, constituindo funções $u_1, \ldots, u_k$ onde $u_i : \Omega \rightarrow [0, 1]$ e $\sum_{i=1}^{k} u_i(x) = 1 \, \forall x \in \Omega$.

Nas partições *fuzzy* não se limita o número de classes às quais os elementos podem pertencer. Nesta abordagem, afirma-se que o elemento x pertence à classe C_i com probabilidade p, se $u_i(x) = p$. Nos casos extremos, pode acontecer que um elemento pertença a uma única classe C_i se $u_j(x) = 0 \, \forall j = 1, \ldots, k (j \neq i)$ e $u_i(x) = 1$; ou pode acontecer que o elemento x pertença a todas as classes se $u_i(x) \neq 0 \;\; \forall i \in \{1, \ldots, k\}$.

Um elemento x é considerado pertencente à classe C_i se $u_i(x)$ apresentar um valor muito perto de 1, ou mais concretamente, escolhe-se C_i como a classe à qual x pertence se $u_i(x) \geq u_j(x)$, $j = 1, \ldots, k, \; j \neq i$.

Em [222] e [223] é definido o algoritmo *Fuzzy c-Means (FCM)* que produz uma partição fuzzy, escolhendo as funções $u_i : \Omega \rightarrow [0, 1]$ tal que $\sum_{i=1}^{k} u_i = 1$ e os centros $r_i \in \mathbb{R}^p$ para $i = 1, \ldots, k$ que minimizam a função

$$\sum_{i=1}^{k} \sum_{j=1}^{n} (u_{ij})^m \parallel x_j - r_i \parallel_2^2 \tag{2.7}$$

onde $\parallel x \parallel_q = (x^T x)^{\frac{1}{q}}$, $k \geq 2$, $m > 1$ e u_{ij} representa a função $u_i(x_j)$, ou seja, representa a força com que o elemento x_j está associado à classe C_i. Definam-se os centros r_i e as funções u_{ij} para $i = 1, \ldots, k$ e $j = 1, \ldots, n$ [223]:

$$r_i = \frac{\sum_{j=1}^{n} (u_{ij})^m x_j^2}{\sum_{j=1}^{n} (u_{ij})^m} \qquad \text{e} \qquad u_{ij} = \frac{\left[\frac{1}{\parallel x_j - v_i \parallel_2^2} \right]^{\frac{1}{m-1}}}{\sum_{i=1}^{k} \left[\frac{1}{\parallel x_j - v_i \parallel_2^2} \right]^{\frac{1}{m-1}}} \tag{2.8}$$

O algoritmo tem como parâmetros de entrada m e k. Prova-se que o algoritmo FCM

converge para um mínimo local [223].

Uma nova abordagem, usando um modelo com funções de pertença proporcionais é apresentada em [170].

2.2 Validação em Análise Classificatória

Existem muitos métodos de classificação ([119], [120], [91], [216] e [60]), não sendo no entanto universais. Dada a diversidade de áreas onde se aplica a análise classificatória, existem algoritmos mais vocacionados para um determinado tipo de dados do que outros, não havendo, no entanto, nenhum algoritmo que seja melhor que todos os outros quaisquer que sejam os elementos a analisar. Basicamente, a análise classificatória é um processo não supervisionado, não se prevendo a existência de qualquer conhecimento prévio da estrutura existente no conjunto de dados. Quase todos os algoritmos de classificação são fortemente dependentes das características do conjunto, assim como dos diferentes valores dos seus parâmetros de entrada. Assim, uma classificação obtida por algum algoritmo de classificação é localmente a melhor possível para os valores dos parâmetros de entrada atribuídos, mas pode não ser localmente a melhor classificação para o conjunto de dados em estudo. Outra questão particularmente preocupante é o facto de qualquer algoritmo de classificação produzir uma classificação no conjunto de dados, mesmo que este seja constituído por dados uniformemente distribuídos sem qualquer estrutura em classes. Outro facto importante é que as classificações do mesmo conjunto de dados, obtidas usando critérios de classificação diferentes, podem diferir sobremaneira umas das outras. Pelo exposto, conclui-se que os algoritmos de classificação podem originar "falsas" classificações, sendo assim essencial que o resultado desses algoritmos seja sujeito a um processo de validação.

Com o objectivo de validar o resultado de uma análise classificatória, o ideal seria optar-se por uma enumeração completa de todas as possíveis partições do conjunto de dados, obtendo-se deste modo uma solução que optimizasse um determinado critério. Esta procura de um óptimo global só é viável para conjunto de dados com poucas dezenas de elementos, situação esta pouco usual em conjuntos de dados reais. A limitação do número de elementos justifica-se pelo número de partições em k classes que é possível estabeler num conjunto de n elementos, que é dado pelo número de *Stirling* de segunda espécie,

$$S(n,k) = \frac{1}{k!}\sum_{i=1}^{k}(-1)^{k-i}\left(\begin{array}{c} k \\ i \end{array}\right) i^n \tag{2.9}$$

À medida que n e k crescem, rapidamente se torna computacionalmente impossível examinar todas as possíveis partições, de modo a identificar uma que optimize o critério de classificação. Verifica-se, para n grande, $S(n,k) \approx \frac{n^k}{k}$, por exemplo $S(19,8) > 1.7 \times 10^{12}$, o que mostra a impraticabilidade desta abordagem mesmo para um conjunto só com 19 elementos e 8 classes.

No entanto, características particulares dos dados ou do critério a optimizar, podem reduzir sensivelmente o número de partições a analisar e tornar possível a obtenção de uma solução óptima. Neste contexto, refira-se o algoritmo de *Fisher* [66] que aplicado a um conjunto de elementos caracterizados por uma variável quantitativa ou qualitativa ordinal, e beneficiando de uma "função de erro" aditiva, permite reduzir consideravelmente o número de partições consideradas, permitindo assim encontrar uma partição óptima de acordo com o critério considerado.

Uma abordagem alternativa consiste em procurar critérios que indiquem se uma determinada solução é ou não satisfatória para o conjunto de dados em análise. No entanto, a informação disponível sobre os dados é frequentemente reduzida ou mesmo nula. Em consequência, surge a ideia por parte de alguns autores, por exemplo [160], de gerar artificialmente dados com uma organização conhecida [161], aplicar um método e verificar se a estrutura existente nos dados é identificada pelos métodos utilizados. Este tipo de testes é útil para testar o desempenho dos algoritmos, no entanto surge o problema da generalização. Será que um algoritmo para o qual se obtém um bom resultado para um determinado conjunto de dados gerado artificialmente, também origina um bom resultado se aplicado a conjuntos de dados, em princípio, com uma estrutura diferente?

Identificando-se alguns problemas no contexto da validação do resultado de uma análise classificatória, têm surgido na literatura alguns trabalhos cujo objectivo principal é validar uma classificação, aumentando a confiança de que esta reflecte realmente a estrutura existente no conjunto de dados.

2.2.1 Diferentes Tipos de Validação

O trabalho realizado em validação é em geral desenvolvido no contexto de um dos três tipos diferentes de testes de validação: externos, internos e relativos:

- Os *testes externos* comparam uma classificação ou parte dela com informação que não é usada para construir a classificação. Neste caso, o resultado do algoritmo é avaliado comparando-o com uma estrutura pré-definida que é imposta no conjunto de dados, reflectindo o estrutura real em classes que se sabe ou que se pensa afectar os elementos do conjunto.

- Os *testes internos* comparam uma classificação ou parte de uma classificação com o conjunto de dados original, usando somente informação obtida a partir do processo de classificação, medindo essencialmente o desvio entre a estrutura gerada pelo algoritmo aplicado e os dados.

- Os *testes relativos* comparam várias estruturas de classificação do mesmo conjunto de dados, resultantes da aplicação do algoritmo com diferentes valores dos parâmetros de entrada ou ao conjunto de dados afectado de pequenas alterações mensuráveis.

A falta de conhecimento da estrutura dos dados, necessária à aplicação de critérios de validação externos, inviabiliza muitas vezes a sua aplicação. No entanto, usando o método de *Monte Carlo* [1], e para alguns conjuntos de dados muito específicos (exibindo propriedades de coesão interna e isolamento externo), em [161] prova-se que existe uma alta correlação entre alguns índices externos e alguns índices internos de validação. Assim, estes índices poderão ser utilizados para avaliar o grau de recuperação de um dado algoritmo da estrutura em classes efectivamente existente nos dados. No referido trabalho, estudaram-se 30 métodos internos de avaliação e dois índices externos de validação que efectivamente medem o grau de recuperação da estrutura existente no conjunto de dados, o índice de *Rand* (Secção 2.5.3) e o índice de *Jaccard* (Secção 2.5.6). Concluiu-se que seis índices internos mostraram ser fiáveis, apresentando uma grande correlação com os índices externos, podendo ser assim considerados índices válidos na identificação do grau de recuperação da estrutura nos dados, substituindo os índices externos.

[1] O método *Monte Carlo* engloba um conjunto de técnicas empíricas com o objectivo de estimar a função de distribuição de índices estatísticos.

No contexto dos testes relativos de validação de classificações, um dos métodos de validar o resultado consiste na avaliação da sua estabilidade ([189], [159], [91] e [198]). O processo engloba a reanálise de uma versão modificada do conjunto de dados, avaliando a extensão das diferenças em relação à classificação original. O conjunto de dados modificado pode ser obtido introduzindo pontos isolados, perturbando as dissemelhanças entre elementos, aumentando a dimensão com variáveis de ruído, ou ainda introduzindo ou removendo alguns elementos ([189] e [159]).

A estabilidade de classificações pode ainda ser avaliada por um método de replicação [33], que segue os seguintes passos:

- O conjunto de dados é dividido de modo aleatório em dois subconjuntos, digamos A e B.
- É aplicada ao subconjunto A uma função apropriada de classificação com o objectivo de obter uma partição daquele em k classes.
- Cada objecto do subconjunto B é atribuído à classe daquela partição que lhe está mais "próxima". Este procedimento proporciona uma partição de B em não mais de k classes.
- A mesma função de partição é aplicada ao subconjunto B para obter uma partição deste em k classes.
- Finalmente são comparadas as duas partições de B, quanto maior for o nível de concordância entre as duas partições, maior a confiança que podemos ter nos resultados.

2.2.2 Modelos Nulos para Avaliação dos Índices de Validação

A avaliação do valor de uma medida de validação de uma estrutura em classes, exige uma comparação com o valor do índice obtido sob um modelo nulo de ausência de estrutura [168]. A ideia base é testar se os elementos do conjunto de dados estão uniformemente estruturados ou não. Esta análise é baseada na *hipótese nula*, H_0, expressa como uma afirmação de uma estrutura uniforme nos dados. A hipótese nula deverá ser testada comparando o valor da estatística de teste, obtida para um conjunto de dados de referência gerados sob a hipótese nula, com o valor da estatística obtida do conjunto de dados a analisar. Assim, será necessário determinar a distribuição do

índice de validação, sob a hipótese nula de ausência de estrutura.

Os modelos nulos sobre os dados podem ser especificados para elementos descritos ou pela matriz dos dados ou pela matriz de dissemelhança, podendo usar ou não alguma informação sobre a estrutura nos dados.

Modelo Uniforme

Neste modelo consideram-se os indivíduos simulados como pontos uniformemente distribuídos pela região A num espaço de dimensão p. Este modelo também pode ser designado por *hipótese de uniformidade* [21] ou *hipótese do posicionamento aleatório* [119].

O principal problema associado a este modelo relaciona-se com a escolha apropriada da região A (janela de amostragem), que influencia sobremaneira os dados gerados. Esta região pode ser considerada como uma hiper-esfera ou um hiper-rectângulo unitário de dimensão p ([227] e [107]). Pode-se ainda considerar a região A como o envelope convexo dos elementos do conjunto de dados ([21] e [168]), no entanto, esta opção de estimação da região A requer pesados requisitos computacionais. Um método alternativo consiste em definir a janela de amostragem como uma hiper-esfera centrada no centro de gravidade dos pontos do conjunto de dados, onde se inclui metade dos seus pontos. Eliminar metade dos pontos não é crucial uma vez que o objectivo desta primeira fase é verificar a existência ou não de uma estrutura em classes no conjunto de dados. Se tal estrutura existir então aplica-se um algoritmo de classificação ao conjunto de dados completo [216].

Modelo Unimodal

No *modelo Unimodal* os indivíduos são simulados supondo que a distribuição conjunta das variáveis que descrevem os elementos, é uma lei multidimensional conjunta Unimodal ([21] e [168]). Assim, o modelo envolve a especificação de uma função de densidade de probabilidade (ou função de probabilidade) para as variáveis que descrevem os elementos. Claramente, também existem dificuldades na justificação da escolha de uma determinada função de densidade para as variáveis. No caso de uma lei Normal multidimensional, salienta-se que os indivíduos que estão próximos do

centro da classe estão mais próximos uns dos outros que os elementos que estão mais afastados do centro da classe, não obstante existir uma única classe.

Modelo de Permutação Aleatória da Matriz de Dados

O *modelo de permutação aleatória da matriz de dados* [91] considera permutações independentes das n entradas, em cada uma das p colunas da matriz de dados. Existem $(n!)^{p-1}$ matrizes essencialmente diferentes e o modelo considera-as equiprováveis. Este modelo ignora a correlação entre variáveis e gera pseudo-objectos dentro de uma região do espaço hiper-rectangular. Este modelo apresenta ainda o inconveniente de que as classes no conjunto de dados original, que são mais homogéneas do que aquelas que são encontradas em muitos conjuntos de pseudo-dados, não têm necessariamente um maior grau de homogeneidade absoluta.

Modelo da Matriz de Dissemelhança Aleatória

Em [148] é proposto um modelo nulo baseado na geração de matrizes de dissemelhanças aleatórias equiprováveis, nas quais os elementos que pertencem ao triângulo inferior são ordenados de modo aleatório, sendo todas as $\frac{n(n-1)}{2}$ possibilidades igualmente prováveis. Este modelo também tem sido referido como *Hipótese do Grafo Aleatório*, no qual cada objecto é representado por um vértice de um grafo e cada arco liga dois vértices se a sua dissemelhança é menor que um valor limite. Uma abordagem alternativa consiste na inserção dos arcos de modo independente e com igual probabilidade.

2.3 Níveis de Validação

A validação em análise classificatória pode ser executada a três níveis. Primeiro deve-se verificar a existência de uma estrutura em classes no conjunto de dados. Se de facto existir tal estrutura, aplica-se um algoritmo de classificação, se não existir, a aplicação do algoritmo de classificação origina uma classificação que não reflecte a estrutura existente no conjunto de dados. O problema de verificar a existência ou não de uma estrutura em classes no conjunto de dados chama-se *tendência de classificação* [216] (Secção 2.3.1).

No procedimento de classificação segue-se a escolha de um "bom" algoritmo de classificação. Por exemplo, John W. Van Ness and Llyd Fisher ([65], [64] e [172]) desenvolveram um formalismo denominado *condições de admissibilidade* (Secção 2.3.2), e que consiste em onze critérios para avaliar algoritmos de classificação. A partir de alguma informação sobre os dados, o método indica quais os critérios de classificação que podem ser relevantes para a análise de um conjunto de dados em particular.

Finalmente deverá proceder-se à *avaliação quantitativa* do resultado do algoritmo de classificação escolhido. Assim, de acordo com o tipo de classificação usado, podemos validar partições ([49], [113], [18], [183], [98] e [95]), partições *fuzzy* ([182], [223] e [222]), hierarquias ([216] e [91])) ou classes ([89], [168] e [91]).

2.3.1 Tendência de Classificação

Uma característica comum à maioria dos algoritmos de classificação é a imposição de uma estrutura em classes no conjunto de dados, mesmo que tal estrutura não exista. É assim fundamental ter um indicador da existência ou não de classes nos dados, antes da aplicação de um algoritmo de classificação. O problema de verificar se o conjunto de dados possui uma estrutura em classes, sem as identificar explicitamente, é denominado *clustering tendency* [216]. Assim, em primeiro lugar deverá ser feita uma verificação da existência de alguma estrutura em classes nos dados, depois, se tal estrutura existir, verifica-se se a classificação obtida pelo algoritmo de classificação seleccionado é ou não fiel aos dados.

Neste contexto, em [91] são referidos alguns métodos que podem ser usados para testar a ausência de uma estrutura em classes no conjunto de dados. Testes do modelo nulo Uniforme têm sido baseados nas seguintes estatísticas de teste: o número de dissemelhanças entre elementos, menores que um determinado valor limite [212]; a maior distância ao vizinho mais próximo de um conjunto de elementos [21]; e comparações entre (i) a distância entre uma posição especificada aleatoriamente e o objecto mais próximo, e (ii) a distância entre este objecto ao seu vizinho mais próximo [109]. Testes do modelo da matriz de dissemelhança aleatória têm sido baseados nas seguintes estatísticas de teste: o número de arcos necessários antes que o grafo seja composto por uma única componente [145]; o número de componentes no grafo quando este é composto por um dado número de arcos ([147] e [171]); e o tamanho das classes quando os elementos são distribuídos em duas classes usando o método de classificação

hierárquica ascendente com o índice do mínimo [47]. Outro teste da homogeneidade do conjunto de dados versus uma estrutura em classes, é desenvolvido sob o modelo de unimodalidade, consistindo na média das distâncias entre todos os pares de elementos do conjunto de dados [21]. Este teste é motivado pela conjectura de que, sob a hipótese de presença de classes, esta média é menor do que sob a hipótese de unimodalidade.

Muitos testes são baseados na árvore de comprimento mínimo no contexto da teoria de grafos. Pode-se representar um grafo no qual os vértices correspondem aos elementos do conjunto, e existe um arco entre quaisquer dois vértices de custo igual à dissemelhança entre eles. Uma árvore de comprimento mínimo é um subgrafo conexo (todos os vértices estão ligados por um caminho ou sequência de outros vértices) sem ciclos e de custo mínimo. O custo de uma árvore deste tipo é usado em [204] para medir a homogeneidade de um conjunto de dados, pelo qual os dados são tanto mais homogéneos quanto maior for o custo da árvore associada. A distribuição dos comprimentos dos arcos de uma árvore de custo mínimo foi estabelecida sob o modelo Normal multivariado [195]. Em [206] é sugerida a introdução de pontos de localização aleatória no conjunto de dados e usar um teste sugerido por [78], baseado no número de arcos na árvore de comprimento mínimo que ligam os pontos originais e os adicionados. Se o conjunto de dados contiver classes, então espera-se que este número seja pequeno. Por outro lado, valores grandes indicam uma organização regular dos elementos do conjunto de dados. Em [78] prova-se que aquele número de arcos segue uma distribuição aproximadamente Normal. Em [200] é descrito um teste pelo qual as árvores de custo mínimo são construídas de tal modo que tenham arcos de comprimento não crescente em todos os caminhos para vértices correspondentes aos centros das classes. Finalmente refira-se uma técnica baseada na *decomposição sequencial* ([216] e [108]) do conjunto de dados, sendo esta uma partição em k conjuntos ou níveis de decomposição, nos quais é importante a ordem pela qual são formados. Os níveis são formados retirando sequencialmente do conjunto de dados os vértices correspondentes ao arco mais longo da árvore de comprimento mínimo definida nos dados. O procedimento de decomposição dá resultados diferentes quando os elementos do conjunto de dados estão organizados em classes, quando estão regularmente espaçados ou quando estão distribuídos de modo uniforme. Baseada nesta observação podem definir-se estatísticas utilizando a informação associada a este procedimento de decomposição.

Outros testes são baseados na procura na multimodalidade no conjunto de dados.

Em [107] foi investigado o número de elementos na menor subclasse de cada classe obtida pelo método de classificação hierárquica ascendente com o índice do mínimo, definindo a estatística de teste como sendo o maior valor que este número pode tomar, de entre todas as classes obtidas por aquele método hierárquico. Neste trabalho foi simulada a distribuição do teste sob os modelos nulos Uniforme e Unimodal. Em [169] e [186] são apresentados testes baseados na comparação entre a quantidade da massa de probabilidade que excede vários valores limite quando existem c modas na distribuição; esta aproximação só é computacionalmente suportável para conjuntos de dados com um reduzido número de elementos.

2.3.2 Condições de Admissibilidade

Considere-se uma colecção de condições de admissibilidade, $A_1, \ldots, A_m$, tendo cada uma delas propriedades desejáveis nos algoritmos de classificação em certas aplicações. Se um algoritmo satisfaz sempre A_j, independentemente do conjunto de dados ao qual é aplicado, então ele é dito ser A_j *admissível*.

Estas condições de admissibilidade constituem um método global e objectivo para comparar algoritmos de classificação e são ([65], [64] e [172]): (1) Space-contracting; (2) Space-conserving; (3) Space-dilating; (4) Positive Admissibility; (5) Metric Admissibility; (6) Monotic Admissibility; (7) Well-structured admissibility; (8) Convex admissibility; (9) Image Admissibility; (10) Connected Admissibility e (11) Proportion Admissibility.

As condições de admissibilidade aplicam-se a algoritmos hierárquicos ascendentes. Sejam C_i e C_j duas classes que são reunidas num dado nível para formarem a classe $C_i \cup C_j$ e seja C_k uma outra qualquer classe. Seja d uma dissemelhança definida entre quaisquer duas classes.

Em cada etapa, após a reunião das classes C_i e C_j, a dissemelhança entre uma outra classe C_k com $k \neq i, j$ e as classes reunidas pode ser: menor ou igual à menor dissemelhança entre C_k e C_i e entre C_k e C_j (*Space-contracting*); pode variar entre a menor e a maior das dissemelhança entre C_k e C_i e entre C_k e C_j (*Space-conserving*); pode ser maior ou igual à maior dissemelhança entre C_k e C_i e entre C_k e C_j (*Space-dilating*) ou pode ser sempre positiva sendo as dissemelhanças entre C_k e C_i e entre

C_k e C_j também positivas (*Positive Admissibility*).

Um algoritmo para o qual é dada uma dissemelhança entre classes, satisfaz a propriedade *Metric Admissibility* se a dissemelhança satisfazer os axiomas de uma métrica, para todos os dados para os quais a dissemelhança original também satisfaz a métrica (ou seja, que satisfaz a desigualdade triangular).

Um algoritmo satisfaz a propriedade *Monotic Admissibility* se ao aplicar-se uma transformação linear a todos os elementos da matriz semelhança ou dissemelhança, o resultado da classificação não é alterado.

Um conjunto de dados é bem estruturado em k grupos se existir uma partição dos dados em k conjuntos, de tal modo que todas as dissemelhanças entre quaisquer dois elementos dentro do mesmo conjunto, é menor que quaisquer dissemelhanças entre dois elementos em conjuntos diferentes. Sendo assim, um algoritmo satisfaz a propriedade *Well-structured admissibility* se identificar correctamente as classes quando aplicado sobre um conjunto bem estruturado.

Dois conjuntos de dados, C_i e C_j, são *Convex Admissible* se os invólucros convexos de cada um deles não se interceptarem.

A propriedade *Image Admissibility* indica que não existe nenhuma outra classificação uniformemente melhor que a dada pelo algoritmo. Ou seja, não existe nenhuma outra partição que tenha alguma dissemelhanças entre elementos pertencentes a uma mesma classe, menor que qualquer dissemelhança entre elementos da mesma classe, na partição obtida pelo algoritmo que goza desta propriedade.

Na aplicação de um algoritmo de classificação hierárquico ascendente a um conjunto Ω, e sempre que duas classes são reunidas, é possível definir uma ligação entre quaisquer pares de elementos, sendo um dos elementos da primeira classe e o outro da segunda classe. Um sistema articulado L_Ω consiste na rede de ligações formada após todos os elementos do conjunto Ω estarem ligados. Uma classificação $C_1, \ldots, C_k$ diz-se *Connected Admissible* se $L_{C_1}, \ldots, L_{C_k}$ são dois a dois disjuntos.

A propriedade *Proportion Admissibility* é desejável quando a geometria das classes

é mais importante que o número de elementos em cada classe. Existem dois tipos de *proportion admissibility*: relativamente a elementos e a classes. Diz-se que um procedimento é *proportion admissible* relativamente a elementos, se a duplicação de um ou vários elementos uma ou mais vezes não alterar as fronteiras das classes. Diz-se que um procedimento é *proportion admissible* relativamente a classes se depois de duplicar cada classe um número arbitrário de vezes, o algoritmo produz classes com as fronteiras inalteradas. Finalmente um algoritmo é *Omission Admissibility* se ao realizar uma nova classificação, após a eliminação de todos os elementos pertencentes a uma classe, o número e a composição das outras classes não for alterado.

As condições de admissibilidade acabadas de descrever são muito exigentes e é extremamente difícil que um algoritmo as satisfaça todas. Não se pretende que todos os algoritmos verifiquem todas as condições de admissibilidade. Pretende-se sim, que para um determinado problema a resolver, se escolham as condições de admissibilidade adequadas ao problema em causa e se procure o algoritmo que as satisfaça.

2.4 Métodos de Comparação de Matrizes que descrevem Classificações Hierárquicas

Alguns métodos de comparação de matrizes podem ser usados no contexto da validação interna de hierarquias. Um dos processos de comparar hierarquias é comparar as matrizes representativas das estruturas hierárquicas. Comecemos com algumas definições.

Definição 2.1 *Seja Ω um conjunto finito, H uma família de partes de Ω (não vazias), H é uma* hierarquia *sobre Ω se:*

1. *$\Omega \in H$, isto é, a parte constituída por todos os elementos está em H.*
2. *$\forall x \in \Omega, \{x\} \in H$, isto é, todas as partes constituídas por um só elemento estão em H.*
3. *Se duas partes $h, h' \in H$ são não disjuntas, $h \cap h' \neq \emptyset$, então uma está necessariamente incluída na outra, $h \subset h'$ ou $h' \subset h$.*

Definição 2.2 *Uma hierarquia indexada é um par (H, f), onde H é uma hierarquia e f uma aplicação, $f : H \mapsto \mathbb{R}^+$ tal que*

1. *Se $h \in H$ verifica-se f(h) =0 se e só se h tem um único elemento.*
2. *Se $h, h' \in H$ e se $h \subset h'$, então $f(h) < f(h')$.*

   Dizemos que uma hierarquia é indexada em sentido lato se, em vez de (2) tivermos

3. *Se $h \subset h'$, então $f(h) \leq f(h')$.*

Partindo de uma árvore hierárquica, podemos construir uma matriz na qual o valor da posição (i, j) corresponde ao índice do nível mais baixo da árvore, a partir do qual os elementos pertencem à mesma classe. Esta métrica é uma distância *ultramétrica*, verificando-se $d(x, y) \leq max\{d(x, z), d(z, y)\}$ para $x, y, z \in \Omega$.

A cada hierarquia indexada podemos fazer corresponder uma (e uma só) distância ultramétrica, e reciprocamente, a cada distância ultramétrica podemos fazer corresponder uma (e uma só) hierarquia indexada (Teorema de *Johnson-Benzécri*). Isto é, existe uma bijecção entre o conjunto das hierarquias indexadas e o conjunto das distâncias ultramétricas sobre um conjunto Ω. A ultramétrica d_u associada a uma hierarquia indexada é definida do seguinte modo:

$d_u(x, y) =$ valor do índice de agregação (altura) da menor classe que contém x e y.

Esta propriedade fornece-nos um modo de calcular a qualidade de uma classificação hierárquica, efectuada a partir de distâncias $d(x, y)$ entre os elementos de Ω. Uma hierarquia é tanto melhor quanto mais semelhantes forem os valores de $d(x, y)$ e $d_u(x, y)$, isto é, quanto menor for a deformação Δ:

$$\Delta = \sum_{x,y \in \Omega} (d(x, y) - d_u(x, y))^2$$

Outra possibilidade será definir uma matriz que represente as distâncias *cladísticas* entre os elementos, estando na posição (i, j) o número de nós (segmentos de linha entre nós) no caminho entre os elementos na estrutura em árvore (não é mais que o número de nós que os separa mais 1). Para algumas árvores, as distâncias cladísticas e ultramétricas podem apresentar elevados graus de correlação. No entanto, os valores ultramétricos são mais fiéis à estrutura hierárquica existente na classificação, sendo as distâncias cladísticas úteis para comparar grafos conexos simples [194]. Outras

possíveis representações de estruturas hierárquicas consistem em matrizes com medidas ultramétricas generalizadas, nomeadamente a métrica de árvores aditivas [114].

Têm sido propostos vários métodos de comparação, ou medição da discrepância, entre duas matrizes, podendo ser usados no contexto dos critérios internos de validação em análise classificatória [194]. Aplicando estes índices à comparação entre uma matriz de dissemelhança e a matriz ultramétrica, obtida após a aplicação de um método de classificação hierárquica, podemos avaliar o grau de acordo que existe entre estas duas matrizes. Quanto maior for este acordo mais eficiente é o método na recuperação da estrutura em classes que organiza o conjunto de dados. Outra abordagem possível consiste na comparação de duas estruturas hierárquicas, sem fazer uso explícito das matrizes de dissemelhança. Por exemplo, refira-se o uso de árvores de *consenso*, que consistem em árvores que contêm os subconjuntos de elementos que são encontrados nas duas classificações a comparar, para desenvolver índices de avaliação do grau de correspondência que existe entre as duas hierarquias ([158], [196] e [92]). Outra classe de métodos para comparar classificações hierárquicas são baseados nas chamadas *transformações métricas*. A distância entre duas classificações pode ser definida como o número mínimo de transformações admissíveis, requeridas para se transformar uma classificação na outra [149].

As matrizes ultramétricas e de dissemelhança obtidas a partir dos dados, podem ainda ser comparadas usando o coeficiente de correlação ultramétrico que é uma medida de concordância, medindo a relação linear que existe entre as matrizes ([194] e [208]). O coeficiente de correlação ultramétrico é também usado em [197] para medir o grau de semelhança entre uma hierarquia (representada pela matriz ultramétrica), e as matrizes de dissemelhança de um conjunto de dados aleatórios (com distribuição multivariada Normal e Uniforme). O estudo foi feito com o objectivo de desenvolver um valor do índice estandardizado, para comparação com resultados obtidos para matrizes de dissemelhança do conjunto de dados a analisar. Pretende-se formular um teste que indique se existe evidência suficiente para se considerar que a relação existente nos dados é hierárquica, e não apenas o que se poderia esperar de uma amostra aleatória de uma única população homogénea.

Baseadas ainda nas matrizes de dissemelhança $P[d_{ij}]$ e nas matrizes ultramétricas $P_u[d^u_{ij}]$ considerem-se as seguintes duas medidas de concordância baseadas em comparações concordantes e discordantes. A primeira consiste numa adaptação da es-

tatística γ de *Goodman and Kruskal* ([87]) para ser usada no contexto da análise classificatória. A medida é dada por [10]:

$$\frac{s(+) - s(-)}{s(+) + s(-)} \tag{2.10}$$

As quantidades $s(+)$ e $s(-)$ representam o número de comparações concordantes e discordantes respectivamente, envolvendo os valores das matrizes $P[d_{ij}]$ e $P_u[d^u_{ij}]$. A comparação entre pares (i, j) e (k, l) pode ser descrita por [91]:

$$d_{ij} R_d d_{kl} \text{ e } d^u_{ij} R_u d^u_{kl}$$

onde os símbolos relacionais verificam $R_d, R_u \in \{<, >\}$. A comparação é dita concordante se $(R_d, R_u) = (<, <)$ ou $(>, >)$, e é dita discordante se $(R_d, R_u) = (<, >)$ ou $(>, <)$. A estatística τ de *Kendall* consiste numa extensão do índice (2.10) de forma a ser normalizado por todos os pares de elementos, inclusivamente os pares de elementos que nem são concordantes nem discordantes, ou seja esta estatística é dada por:

$$\frac{s(+) - s(-)}{n(n-1)/2} \tag{2.11}$$

Uma metodologia de validação de classificação hierárquica é apresentada por [210]. Neste trabalho, a geração de árvores, tendo em conta a sua topologia, os seus níveis e as etiquetas das folhas, é usada para obter matrizes ultramétricas. Estas matrizes são submetidas a diferentes critérios de classificação hierárquica ascendente, e consequente comparação entre a estrutura inicial existente nos dados e a estrutura hierárquica obtida para cada critério, com vista a medir a capacidade de cada método em recuperar a estrutura existente.

2.5 Medidas de Comparação entre Duas Partições

A validação externa e relativa das partições obtidas por métodos de classificação, exigem a aplicação de medidas que avaliem o grau de semelhança entre partições. Em alguns casos, é necessário realizar uma comparação entre as partições obtidas e a partição de referência (validação externa). Noutros casos, pretende-se comparar diferentes partições obtidas para o mesmo conjunto de dados, resultantes, ou da aplicação de diferentes algoritmos de classificação, ou da aplicação do mesmo algoritmo

de classificação para diferentes valores dos seus parâmetros de entrada (validação relativa). Designemos por U e V as partições a comparar. A partição U divide a população em R classes, $u_1, u_2, \ldots, u_R$, enquanto que a partição V divide a população em C classes, $v_1, v_2, \ldots, v_C$. Algumas das medidas de comparação mais conhecidas são baseadas em tabelas de contingência e pares de elementos concordantes. Comecemos esta secção por definir cada um destes conceitos.

2.5.1 Tabela de Contingência. Pares de Elementos Concordantes.

Uma tabela de contingência representa a distribuição dos elementos pelas classes de U e de V. Cada elemento n_{ij} (ρ_{ij}) da tabela representa o número (proporção) de elementos classificados em u_i na primeira partição e em v_j na segunda.

A tabela pode então ser definida em relação à frequências absolutas (proporções):

		partição V				
	Classes	v_1	v_2	$\ldots$	v_C	Totais
partição U	u_1	$n_{11}(\rho_{11})$	$n_{12}(\rho_{12})$	$\ldots$	$n_{1\beta}(\rho_{1\beta})$	$n_{1.}(\rho_{1.})$
	u_2	$n_{21}(\rho_{21})$	$n_{22}(\rho_{22})$	$\ldots$	$n_{2C}(\rho_{2C})$	$n_{2.}(\rho_{2.})$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	u_R	$n_{R1}(\rho_{R1})$	$n_{R2}(\rho_{R2})$	$\ldots$	$n_{RC}(\rho_{RC})$	$n_{R.}(\rho_{R.})$
	Totais	$n_{.1}(\rho_{.1})$	$n_{.2}(\rho_{.2})$	$\ldots$	$n_{.C}(\rho_{.C})$	$n_{..} = n(\rho_{..} = 1)$

Onde $n_{i.}(\rho_{i.})$, $(i = 1, \ldots, R)$ representa o número (proporção) de elementos da população classificada em u_i, e $n_{.j}(\rho_{.j})$ $(j = 1, \ldots, C)$ representa o número (proporção) de elementos da população classificada em v_j. Na tabela de contingência, a linha i representa a distribuição dos elementos da classe u_i por todas as classes da partição V, e a coluna j representa a distribuição dos elementos da classe v_j por todas as classes da partição U.

Muitas medidas são calculadas em função do número de pares de valores de Ω, $\{x_i, x_j\}$ que são classificados do mesmo modo nas duas partições U e V. Designemos quatro tipos de atribuições diferentes que podem estar associadas ao par $\{x_i, x_j\}$ [113]:

- Tipo I: O par é atribuído à mesma classe em U e em V. O número de casos

deste tipo representa-se por a.

- Tipo II: O par é atribuído a classes diferentes em U e em V. O número de casos deste tipo representa-se por b.
- Tipo III: O par é atribuído a classes diferentes em U mas na mesma classe em V. O número de casos deste tipo representa-se por c.
- Tipo IV: O par é atribuído à mesma classe em U mas em classes diferentes em V. O número de casos deste tipo representa-se por d.

Os pares de elementos de Tipo I e Tipo II dizem-se *pares concordantes*, os restantes dizem-se *pares discordantes*.

Defina-se por $A = a + b$ o número de todos os pares concordantes (pares do tipo I e II) e por $D = c + d$ o número de pares discordantes (pares de tipo III e IV). Então tem-se [113]:

$$A = \binom{n}{2} + \sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}^2 - \frac{1}{2}\left(\sum_{i=1}^{R} n_{i.}^2 + \sum_{j=1}^{C} n_{.j}^2\right) \tag{2.12}$$

e

$$A + D = \binom{n}{2} \Longleftrightarrow D = \frac{1}{2}\left(\sum_{i=1}^{R} n_{i.}^2 + \sum_{j=1}^{C} n_{.j}^2\right) - \sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}^2 \tag{2.13}$$

2.5.2 Índice *Kappa de Cohen*

Um dos primeiros índices que surgiram na literatura usando pares de elementos concordantes foi o índice *Kappa de Cohen*, aplicado no caso particular em que $R = C$ (mesmo número de classes nas duas partições), e é dado por ([44] e [116]):

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{2.14}$$

onde

$$P_o = \sum_{i=1}^{R} \frac{n_{ii}}{n}$$

é a proporção observada de pares concordantes e

$$P_e = \sum_{i=1}^{R} \frac{n_{i.} n_{.i}}{n^2}$$

é a proporção esperada de pares concordantes devido ao acaso e sob a hipótese de independência entre partições.

2.5.3 Índice de Rand

O índice de *Rand* permite comparar duas partições com número de classes não necessariamente iguais. Basicamente, este índice baseia-se no número de pares de elementos que foram atribuídos da mesma maneira em cada uma das partições, ou seja, baseia-se no número de pares de elementos concordantes (tipo I e II). Assim, seguindo a notação da secção anterior, o índice de *Rand*, designado por IR é dado por [189]:

$$IR = \frac{a+b}{a+b+c+d} = \frac{A}{\binom{n}{2}}$$

De um modo mais pormenorizado tem-se:

$$IR = \frac{\binom{n}{2} + \sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}^2 - \frac{1}{2}\left[\sum_{i=1}^{R} n_{i.}^2 + \sum_{j=1}^{C} n_{.j}^2\right]}{\binom{n}{2}} \tag{2.15}$$

Verifica-se $0 \leq IR \leq 1$, tomando o valor 0 quando as duas partições não têm qualquer semelhança (ou seja, quando uma partição é constituída por uma só classe com todos os elementos, e a outra é constituída por n classes com 1 elemento cada) e o valor 1 quando o acordo entre as duas partições é completo. Em [189] foi definida de modo assimptótico a distribuição do índice descrita pela média, pelo desvio padrão e pela percentagem de completo acordo, usando técnicas de simulação de *Monte Carlo*. Outros índices baseados em pares de elementos concordantes são propostos em [219] e [69].

Um dos problemas deste índice é a não definição do valor que assume, se o acordo entre as duas partições for devido ao acaso. Com o objectivo de colmatar esta lacuna, outros autores se debruçaram sobre este índice e deram um passo em frente no seu tratamento probabilístico ([35], [116], [38], [167], [113] e [132]).

O valor esperado do número de pares concordantes depende das entradas da tabela de contingência n_{ij}, cuja distribuição depende, por sua vez, de os subtotais marginais da tabela serem fixos (modelo hipergeométrico) ou variáveis (modelo multinomial).

Usando o modelo hipergeométrico em [35] são deduzidas as fórmulas para a média e a variância do número de pares concordantes (A). A hipótese nula de acordo devido ao acaso pode ser testada usando a estatística de *Brennan & Light* [35]: $Z_A = \dfrac{A - E(A)}{\sqrt{var(A)}}$ e assumindo que, para grandes valores de R, C e n, segue uma distribuição Normal estandardizada. Os autores fizeram um estudo empírico para testar a validade da aproximação à Normal tomada pela estatística apresentada, usando valores pequenos de R, C e n e obtiveram uma boa aproximação, mostrando ainda que a validade da aproximação depende mais dos valores de R e de C do que de n.

Em [116] é apresentada uma generalização do estudo apresentado em [35], mais concretamente é obtida uma função monótona da estatística de *Brennan & Light*, baseada na generalização de um coeficiente de correlação, acompanhada de uma interpretação probabilística, permitindo uma simplificação do cálculo daquela estatística. Considerando somas marginais fixas e variáveis, são deduzidas as fórmulas para a média e a variância do número de pares concordantes. No mesmo contexto, é sugerida uma estatística com uma distribuição assimptoticamente Normal de uma amostra de grande dimensão, para testar a hipótese nula de que o acordo existente nos pares concordantes seja devido ao acaso.

Em [38] é referido que a distribuição Normal para o número de pares concordantes definido em [35] é inadequada, particularmente quando uma ou as duas partições têm um número muito aproximado de elementos em cada classe. Neste trabalho prova-se que se as margens da tabela de contingência são constantes, A é claramente uma função linear da estatística χ^2 para testar a associação da tabela e espera-se que seja aproximadamente uma função linear de uma distribuição do qui-quadrado com $(R-1)(C-1)$ graus de liberdade. Isto sugere que a distribuição de A deverá estar longe da distribuição Normal se R e C forem pequenos mesmo que n seja grande. Usando o método dos momentos fez-se uma aproximação da distribuição de A por uma função linear da distribuição do qui-quadrado, defendendo que a aproximação do qui-quadrado da estatística A é melhor do que a aproximação à Normal.

2.5.4 Índice de Rand Ajustado de Morey e Agresti

Nesta secção é proposto um ajuste do índice de *Rand* permitindo ao utilizador comparar duas partições diferentes, corrigindo o índice de modo a que tome em conta o contributo do acaso para qualquer semelhança verificada.

O *Índice de Rand Ajustado de Morey e Agresti* (IRA) consiste num ajustamento do índice de *Rand* para o acaso e é definido por [167]:

$$IRA = \frac{\sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}^2 - \frac{\sum_{i=1}^{R} n_{i.}^2 \sum_{j=1}^{C} n_{.j}^2}{n^2}}{\frac{1}{2}\left[\sum_{i=1}^{R} n_{i.}^2 + \sum_{j=1}^{C} n_{.j}^2\right] - \frac{\sum_{i=1}^{R} n_{i.}^2 \sum_{j=1}^{C} n_{.j}^2}{n^2}} \tag{2.16}$$

Verifica-se $-1 \leq IRA \leq 1$, o índice toma o valor é 1 quando as duas partições são idênticas, toma o valor 0 quando o acordo verificado entre as duas partições se deve ao acaso e toma valores negativos quando o grau de semelhança entre as duas partições é menor que o valor esperado por uma atribuição ao acaso.

Se se classificar uma amostra aleatória com número fixo de classes, então a estatística IRA segue uma distribuição aproximadamente Normal. A variância assimptótica para a estatística IRA é calculada usando o método delta [88].

2.5.5 Índice de Rand Corrigido de Hubert e Arabie

A dedução apresentada em [167] (e usada em [164]) é contestada por *Hubert & Arabie* em [113] por ser desenvolvida com base num pressuposto falso, o qual advogava que o valor esperado do quadrado de uma variável aleatória é igual ao quadrado do valor esperado dessa variável.

Vejamos então como em [113] é feita a correcção da estatística de *Rand* para o acaso, associando-lhe uma interpretação probabilística, ou seja, tomando em conta

o contributo do acaso no valor do índice que terá de ser um valor constante (ex. zero).

Sob o pressuposto hipergeométrico para as somas marginais da tabela de contingência, mostra-se que o valor esperado do índice de *Rand* é dado pela seguinte expressão:

$$E(IR) = 1 + 2\frac{\sum_{i=1}^{R}\binom{n_{i.}}{2}\sum_{j=1}^{C}\binom{n_{.j}}{2}}{\binom{n}{2}^2} - \frac{\sum_{i=1}^{R}\binom{n_{i.}}{2} + \sum_{j=1}^{C}\binom{n_{.j}}{2}}{\binom{n}{2}} \tag{2.17}$$

Usando a fórmula geral de um índice I corrigido para o acaso:

$$\frac{I - E(I)}{I_{max} - E(I)} \tag{2.18}$$

que é limitado por 1 e toma o valor zero quando o índice iguala o seu valor esperado, o *Índice de Rand Corrigido de Hubert e Arabie* (IRC) é dado por [113]:

$$IRC = \frac{\sum_{i=1}^{R}\sum_{j=1}^{C}\binom{n_{ij}}{2} - \binom{n}{2}^{-1}\sum_{i=1}^{R}\binom{n_{i.}}{2}\sum_{j=1}^{C}\binom{n_{.j}}{2}}{\frac{1}{2}\left[\sum_{i=1}^{R}\binom{n_{i.}}{2} + \sum_{j=1}^{C}\binom{n_{.j}}{2}\right] - \binom{n}{2}^{-1}\sum_{i=1}^{R}\binom{n_{i.}}{2}\sum_{j=1}^{C}\binom{n_{.j}}{2}} \tag{2.19}$$

Este índice toma valores em $[-1, 1]$, onde o valor 1 indica um perfeito acordo entre as duas partições, enquanto que valores próximos de 0 correspondem a um acordo entre as partições devido ao acaso.

Este índice vai ser usado ao longo da dissertação sempre que for necessário comparar duas partições.

2.5.6 Índice de *Jaccard*

O índice de *Jaccard* é dado por

$$J = \frac{a}{a + c + d} \tag{2.20}$$

Este índice ignora os casos de tipo II, ou seja, não considera os casos em que os pares de elementos são classificados em classes diferentes em ambas as partições. Em [160], na realização de algumas experiências, verifica-se que o índice de *Rand* e de *Jaccard* apresentaram uma elevada correlação (0.937), mostrando que de facto, os pares de tipo II podem não ter uma grande influência na comparação de partições.

2.5.7 Índice de Fowlkes & Mallows

Em [69] é apresentada uma medida numérica do grau de semelhança entre classificações hierárquicas representada por B_k, sendo calculada a partir das partições obtidas pelo corte das árvores representativas das classificações hierárquicas, num nível tal que origine k classes em cada uma. Deste modo, e tal como foi notado em [173] e [219], esta medida é aplicada a partições sem se ter em conta a estrutura hierárquica em cada uma das classificações, por exemplo o nível onde se cortou cada uma das árvores é ignorado. Para este índice é calculada a média e a variância sob o pressuposto de que as somas marginais da matriz $[n_{ij}]$ são constantes, ou seja, considera-se constante o número de elementos em cada uma das classes de cada uma das classificações. Através de algumas experiências baseadas no método *Monte Carlo*, compara-se B_k com o *Índice de Baker* [9] e com o *Índice de Rand* [189]. Nesta secção vamos considerar o caso particular em que na tabela de contingência se tem $R = S = k$ (apesar de esta restrição não ser necessária segundo [219]). A medida proposta é dada por [69]:

$$B_k = \frac{T_k}{\sqrt{P_k\, Q_k}} \tag{2.21}$$

onde

$$T_k = \sum_{i=1}^{k}\sum_{j=1}^{k}\binom{n_{ij}}{2} = \sum_{i=1}^{k}\sum_{j=1}^{k} n_{ij}^2 - n \tag{2.22}$$

$$P_k = \sum_{i=1}^{k}\binom{n_{i.}}{2} = \sum_{i=1}^{k} n_{i.}^2 - n \tag{2.23}$$

$$Q_k = \sum_{j=1}^{k}\binom{n_{.j}}{2} = \sum_{j=1}^{k} n_{.j}^2 - n \tag{2.24}$$

$$n_{i.} = \sum_{j=1}^{k} n_{ij} \quad ; \quad n_{.j} = \sum_{i=1}^{k} n_{ij} \quad ; \quad n_{..} = n = \sum_{i=1}^{k}\sum_{j=1}^{k} n_{ij} \tag{2.25}$$

Verifica-se $0 \leq B_k \leq 1$, existindo uma correspondência directa entre valores altos do índice e um acordo entre as duas partições a serem comparadas. Sob o pressuposto hipergeométrico para as somas marginais da tabela de contingência, os autores deduzem as expressões para a média e a variância de B_k. Neste trabalho é deduzida outra fórmula para o índice de Rand baseada nas expressões (2.22), (2.23) e (2.24) e usando o mesmo pressuposto de margens fixas, são deduzidas a média e a variância para este índice.

O índice de *Fowlkes & Mallows* também pode ser dado por [163]:

$$B_k = \frac{a}{\sqrt{(a+b)(a+c)}} \tag{2.26}$$

2.5.8 Índice de Youness & Saporta

Mais recentemente em [225] foi proposto um novo índice de comparação de duas partições do mesmo conjunto de dados, baseado no teste de *Mac Nemar*. Este é um teste não paramétrico frequentemente usado para verificar a igualdade de duas proporções em amostras encaixadas. Por exemplo, represente-se por a' o número de indivíduos que mantêm a mesma opinião favorável, antes e depois de uma campanha; represente-se por b' o número de indivíduos que mantêm a mesma opinião desfavorável, antes e depois de uma campanha; finalmente, considere-se c' e d' a frequência daqueles que mudaram de opinião no decurso da campanha. A estatística de teste correspondente à hipótese nula de que as mudanças de opinião são equiprováveis é dada por [225]:

$$Mc = \frac{d' - c'}{\sqrt{c' + d'}} \tag{2.27}$$

que segue uma distribuição aproximadamente Normal de média zero e desvio um, para n grande.

Usando o teste para o conjunto de pares de objectos, obtém-se uma nova medida de comparação entre duas partições [225]:

$$Mc = \frac{\sum_{i=1}^{R} n_{i.}^2 - \sum_{j=1}^{C} n_{.j}^2}{2\sqrt{\frac{1}{2}\left(\sum_{i=1}^{R} n_{i.}^2 - \sum_{j=1}^{C} n_{.j}^2\right) - \sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}^2}} \tag{2.28}$$

2.5.9 Comparação entre Alguns Critérios Externos

Já foram feitos alguns trabalhos de comparação entre alguns índices apresentadas nesta secção, entre eles refira-se um trabalho apresentado em [164] no qual usando técnicas de *Monte Carlo* se comparam os índices: índice de *Rand* (IR), índice de *Rand* Ajustado de *Morey e Agresti* (IRA), índice de *Jaccard* e índice de *Fowlkes & Mallows*. Como resultado das experiências concluiu-se que os índices IRA e de *Jaccard* apresentaram um melhor desempenho na avaliação da capacidade de recuperação da estrutura existente nos dados por um método de classificação hierárquica. No entanto, neste trabalho os índices só foram aplicados a partições com o mesmo número de classes. Em [163], os quatro índices referidos no trabalho anterior e ainda o índice de Rand Corrigido de *Hubert & Arabie* (IRC) foram comparados usando partições com diferente número de classes. Os índices são usados na comparação do resultado obtido pela aplicação de um algoritmo hierárquico ao conjunto de dados, com a partição que indica a estrutura em classes que se sabe existir no conjunto de dados.

O estudo apresentado foi efectuado comparando uma partição com um número de classes fixo representando a estrutura que se sabe existir no conjunto de dados, com diferentes níveis de uma hierarquia, ou seja, com partições com diferente número de classes. O resultado do estudo mostrou que os índices IRA e IRC apresentaram um bom comportamento para o caso nulo de ausência de classes, concluindo-se, no entanto, que o *Índice de Rand Ajustado de Hubert e Arabie* foi mais eficiente.

Capítulo 3

Escolha do Número de Classes

Sendo a análise classificatória um processo não supervisionado, não é suposto existir algum conhecimento sobre a estrutura dos dados, capaz de fornecer informação útil para a identificação do número de classes. Assim, um dos problemas mais persistentes no contexto da análise classificatória, consiste no desconhecimento *a priori* do número de classes existente no conjunto de dados. O problema é muito pertinente uma vez que o número de classes é um dos parâmetros de entrada de um elevado número de algoritmos de classificação, condicionando sobremaneira a eficácia dos métodos. Assim, torna-se muito importante toda a investigação realizada nesta área.

O processo que orienta a escolha do número de classes existente no conjunto de dados, é diferenciado de acordo com o tipo de algoritmo de classificação usado. Quanto aos algoritmos de partição, à parte de algumas excepções, a investigação tem consistido no desenvolvimento de alguns índices de validação que, associados a algoritmos de classificação e aos seus parâmetros de entrada, pretendem indicar o número de classes existente no conjunto de dados. Pretende-se seleccionar o algoritmo e a combinação dos valores dos seus parâmetros de entrada, que correspondem à partição que melhor descreve a estrutura em classes dos dados. Seguindo esta abordagem, os algoritmos de classificação são executados para diferentes valores dos seus parâmetros de entrada, obtendo-se uma partição para cada combinação de valores, para a qual é calculado um índice de validação. Escolhe-se a partição em k classes, aquela que obedece ao critério estabelecido pelo índice de validação. Se o algoritmo de classificação tiver como único parâmetro de entrada o número de classes k, faz-se variar o seu valor, em geral de 2 até um número máximo k_{max}. Um dos problemas que surgem neste contexto, diz respeito ao valor de k_{max} a escolher. Na literatura não existem guias de acção para

uma boa escolha deste valor; no entanto, se não existir qualquer informação àcerca do número de classes presente no conjunto de dados, pode considerar-se $k_{max} \leq \sqrt{n}$ [182].

Esta estratégia funciona se o índice não for tendencioso com o número de classes. Infelizmente muitos índices de validação conhecidos apresentam um aumento (diminuição) do seu valor, à medida que o número de classes aumenta. Sendo assim, localizar um mínimo (máximo) local indicador do valor de k, pode não identificar uma boa partição. Para índices que apresentem esta indesejável tendência, a alternativa é procurar o valor de k para o qual ocorre localmente uma mudança significativa do valor do índice de validação, identificando-se graficamente como um "cotovelo". Uma outra fonte de complicações, associadas a muitos dos índices neste contexto, é que os seus comportamentos dependem de muitos outros factores tais como o número e a dimensionalidade dos elementos no conjunto de dados [216].

Nas excepções à abordagem anterior refiram-se os métodos baseados numa mistura de modelos nos quais se assume que os dados são gerados por um modelo misto, no qual cada componente é interpretado como uma classe, e pressupõe-se que cada elemento do conjunto de dados foi gerado por uma e uma só das classes ou componentes. Neste tipo de abordagem, dado o conjunto de dados no qual não é conhecida a componente origem de cada elemento, são inferidos os parâmetros das funções de densidade (as classes) ([11] e [209]). O número de classes que melhor reflectem a estrutura do conjunto de dados, pode ser estimado usando modelos probabilísticos baseados em *testes de Monte Carlo com validação cruzada* [207]; ou pode ser inferido supondo uma mistura de modelos gaussianos nos dados e pela detecção de um "cotovelo" na curva de máxima verosimilhança [19]; ou ainda usando uma aproximação Bayesiana, pela qual o problema da escolha do número de classes é resolvido em simultâneo com a escolha do melhor modelo ([70] e [11]). Em [43] é apresentado o método *AUTOCLASS* que é um algoritmo baseado no modelo clássico de mistura de componentes, suplementado por uma método *Bayesiano* para determinar o número de classes. Neste método não se considera que cada elemento pertence a uma única componente ou classe, mas são calculadas as probabilidades de um elemento pertencer a cada uma das classes. Em [215] são apresentados novos critérios de classificação para serem usados no caso em que se pode supor uma mistura de distribuições normais multivariadas no conjunto de dados. Os critérios baseiam-se em abordagens *Bayesianas* e de máxima verosimilhança, correspondentes a diferentes pressupostos sobre as matrizes de covariância da mistura de componentes. Os métodos de reamostragem, habitualmente utilizados no

contexto da análise discriminante, podem também ser usados para estimar o número de classes em análise classificatória [77].

Têm surgido na literatura muitos índices de validação considerados como medidas da qualidade de uma classificação, pretendendo identificar estruturas que reflectem melhor a organização dos dados em classes, quando tal estrutura existe. No entanto, é muito difícil definir índices de validação que sejam completamente fiáveis para qualquer que seja o conjunto de dados. O melhor que se consegue é a definição de alguns índices de validação que apresentam uma elevada fiabilidade para conjuntos de dados com determinadas características. É difícil definir um índice de validação que seja fiável para todos os conjuntos de dados. Tal como referido em [182]: *"no matter how good your index is, there is a data set out there waiting to trick it (and you)".*

Segue-se a descrição de alguns índices de validação que têm como tarefa principal identificar o número de classes existente no conjunto de dados. Estes índices inserem-se no contexto da validação relativa do resultado da aplicação de um algoritmo de classificação, uma vez que consiste na comparação de diferentes partições de um mesmo conjunto de dados, resultantes da aplicação do mesmo algoritmo com diferentes valores para os seus parâmetros de entrada.

3.1 Índice de Validação SD

O *Índice de validação SD* é proposto em [104] e a sua definição é baseada nos conceitos de dispersão média das classes e separação total entre classes, definidas usando dissemelhanças intra-classes e entre classes. Basicamente, este índice selecciona a partição cuja densidade de pontos dentro das classes é muito maior do que a densidade de pontos entre classes. Assim, o índice pretende identificar classes compactas e isoladas. Consideremos $\Omega = \{x_1, \ldots, x_n\}$ o conjunto de n elementos descritos por t atributos numéricos, e $C_1, \ldots, C_k$ uma partição em k classes de $n_1, \ldots, n_k$ elementos e $r_1, \ldots, r_k$ representantes, respectivamente. Represente-se por $\zeta(\Omega)$ a dispersão do conjunto de dados, definida para a $t^{\text{ésima}}$ coordenada, por $\zeta_t(\Omega) = \frac{1}{n}\sum_{j=1}^{n}(x_{jt} - \bar{x}_t)^2$ onde $\bar{x}_t$ é a $t^{\text{ésima}}$ coordenada de $\bar{X} = \frac{1}{n}\sum_{j=1}^{n} x_j$, $\forall x_j \in \Omega$. Represente-se por $\zeta(C_i)$ a

dispersão da classe C_i e defina-se, para a $t^{\text{ésima}}$ coordenada, $\zeta_t(C_i) = \frac{1}{n_i} \sum_{j \in A_i} (x_{jt} - r_{it})^2$ onde r_i é o representante da classe C_i e A_i é o conjunto dos índices dos elementos pertencentes à classe C_i.

O índice tem uma ordem de grandeza linear, e é definido por:

$$SD(k) = \gamma Scat(k) + Dis(k) \tag{3.1}$$

onde $Scat(k) = \frac{1}{k} \sum_{i=1}^{k} \frac{\| \zeta(C_i) \|_2}{\| \zeta(\Omega) \|_2}$ ([1]) e $Dis(k)$ é a separação total entre classes, dada por

$$Dis(k) = \frac{D_{max}}{D_{min}} \sum_{i=1}^{k} \left(\sum_{j=1}^{k} \|r_i - r_j\|_2 \right)^{-1} \tag{3.2}$$

onde $D_{max} = max\{\|r_i - r_j\|_2\} \forall i, j \in \{1, 2, \ldots, k\}$ e $D_{min} = min\{\|r_i - r_j\|_2\} \forall i, j \in \{1, 2, \ldots, k\}$ são respectivamente a distância máxima e mínima entre representantes das classes. O factor γ é um factor de balanço e é igual a $Dis(k_{max})$ quando k_{max} é o número máximo de classes. A influência do número máximo de classes relacionado com o factor de balanço é discutido em [102], sendo provado que o valor indicativo do número de classes é obtido quase independentemente do valor máximo de classes. O número de classes presentes no conjunto de dados é aquele que minimiza o valor do índice.

3.2 Índice de *Davies-Bouldin* (DBI)

O índice de *Davies-Bouldin* é baseado na ideia de que uma boa partição é aquela para a qual as medidas de separação entre classes e a densidade das classes apresentam valores altos. O índice é uma função da razão entre a soma da dispersão intra-classes e a separação entre-classes e é dado por [49]:

$$\bar{R} = \frac{1}{k} \sum_{i=1}^{k} max_{\{i \neq j\}} \{R_{ij}\} \tag{3.3}$$

[1] $\| x \|_q = (x^T x)^{\frac{1}{q}}$

onde

$$R_{ij} = R(s_i^q, s_j^q, d_{ij}) = \frac{s_i^q + s_j^q}{d_{ij}} \tag{3.4}$$

sendo $s_i^q = \frac{1}{n_i} \sum_{j \in A_i} \parallel x_j - r_i \parallel_q$ e $s_j^q = \frac{1}{n_j} \sum_{i \in A_j} \parallel x_i - r_j \parallel_q$ as dispersões das classes C_i e C_j respectivamente, r_i e r_j e A_i e A_j são respectivamente os centros de gravidade e os conjuntos dos índices dos elementos das classes C_i e C_j. Finalmente d_{ij} é a dissemelhança entre os centros das classes C_i e C_j.

Uma vez que o objectivo é minimizar a dispersão dentro de classes e maximizar a separação entre classes, o número de classes k que minimiza $\bar{R}$, é tomado como o valor indicativo do número de classes existente no conjunto de dados.

3.3 Índice de *Dunn* (DI)

O índice de *Dunn* tenta identificar partições de classes compactas e isoladas. O número de classes que maximiza o valor do índice é tomado como o valor de k, adequado ao conjunto de dados. O índice é dado por [18]:

$$DI = \min_{\{i=1,\ldots,k-1\}} \left\{ \min_{\{j=i+1,\ldots,k\}} \left\{ \frac{\delta(C_i, C_j)}{\max_{\{r=1,\ldots,k\}}\{\Delta(C_r)\}} \right\} \right\} \tag{3.5}$$

onde $\delta(C_i, C_j)$ é a dissemelhança entre as classes C_i e C_j, dada por exemplo, pela média das dissemelhanças entre pares de elementos de cada uma das classes. O valor $\Delta(C_r)$ representa o diâmetro da classe C_r, sendo dado, por exemplo, pela dissemelhança máxima entre pares de elementos da classe.

Se o conjunto de dados contém classes compactas e bem separadas, então espera-se que a dissemelhança entre as classes seja grande e o diâmetro das classes seja pequeno. Sendo assim, e baseando-nos nas definições do índice de *Dunn*, pode-se concluir que grandes valores do índice indicam a presença de classes compactas e bem separadas.

O desempenho do índice de *Dunn* depende sobremaneira das definições da dissemelhança entre classes $\delta(\cdot,\cdot)$ e do diâmetro das classes $\Delta(\cdot)$. Em [18] foi feito um estudo de comparação entre várias definições para estas duas medidas. Para a dissemelhança entre classes o autor definiu seis medidas diferentes: o mínimo entre as dissemelhanças de pares de elementos pertencentes a classes diferentes (δ_1); o máximo entre as dissemelhanças de pares de elementos pertencentes a classes diferentes (δ_2); a média entre as dissemelhanças de pares de elementos pertencentes a classes diferentes (δ_3); a dissemelhança entre os centros das classes (δ_4); a média das dissemelhanças entre os elementos das duas classes e os respectivos centros (δ_5); e ainda a métrica de *Hausdorff* (δ_6) definida por:

$$\delta_6(S,T) \;=\; \delta_{Hausdorff}(S,T) \;=\; max\{\delta(S,T),\delta(T,S)\} \tag{3.6}$$

onde

$$\begin{aligned} \delta(S,T) &= max_{\{x\in S\}}\{min_{\{y\in T\}}\{d(x,y)\}\} \\ \delta(T,S) &= max_{\{y\in T\}}\{min_{\{x\in S\}}\{d(x,y)\}\} \end{aligned} \tag{3.7}$$

sendo esta medida menos sensível à presença de pontos isolados.

Foram definidas três medidas para o diâmetro de uma classe: o máximo das dissemelhanças entre pares de elementos da classe (Δ_1); a média das dissemelhanças entre pares de elementos da classe (Δ_2); e a média das dissemelhanças entre os elementos da classe e o centro da mesma (Δ_3).

Combinando estas medidas é possível definir dezoito variações do índice de *Dunn*. Das experiências realizadas [18] concluiu-se que os índices de *Dunn* que usam a medida δ_1 ou Δ_1, apresentam um pior desempenho devido à grande sensibilidade destas medidas relativamente a pontos isolados. Este facto é consequência de os cálculos englobarem só dois dos elementos da classe, e não todos, como acontece para algumas outras medidas. As medidas que conjugam Δ_1 com δ_3, δ_4 ou δ_5; e Δ_3 com δ_3 ou δ_5 apresentam bons desempenhos. No estudo referido, as medidas δ_3 e δ_5 foram eficazes, sendo referido que outros estudos mostram o bom desempenho de δ_6.

3.4 Generalização dos Índices *Davies-Bouldin* e de *Dunn* usando Conceitos da Teoria dos Grafos

Em [183] são apresentadas formas generalizadas para os índices de *Davies-Bouldin* (BDI) e de *Dunn* (DI), utilizando conceitos da teoria de grafos, de modo a abranger classes não esféricas e serem robustos na presença de ponto isolados.

Considere-se uma correspondência directa entre os vértices do grafo e os elementos do conjunto de dados. Os grafos Árvore de Comprimento Mínimo (*Minimal Spanning Tree*) (ACM), Grafo de Vizinhança Relativa (*Relative Neighborhood Graph*) (GVR) e Grafo Gabriel (*Gabriel Graph*) (GG) são frequentemente usados no contexto da análise classificatória, e são nesta secção aplicados à construção de índices de validação.

Definamos resumidamente cada um deste grafos. Consideremos um grafo representativo dos dados, no qual os arcos estão ponderados pela dissemelhança entre os vértices, representando o custo do arco. Quaisquer dois vértices estão ligados por um caminho ou sequência de arcos, cujo custo é igual ao custo máximo de entre todos os arcos presentes no caminho. Um grafo ACM é um grafo deste tipo sem ciclos e de custo mínimo. Num grafo GVR, dois vértices x_i e x_j estão ligados se se verificar $d(x_i, x_j) \leq max\{d(x_i, x_k), d(x_j, x_k),\ \forall k,\ k \neq i, k \neq j\}$ onde $d(x_i, x_j)$ é a distância euclideana entre x_i e x_j; ou seja, se nenhum outro vértice do conjunto pertence à intersecção das duas bolas de raio $d(x_i, x_j)$ e de centros x_i e x_j. Finalmente num grafo GG, dois vértices x_i e x_j estão ligados se se verificar $d^2(x_i, x_j) < d^2(x_i, x_k) + d^2(x_j, x_k),\ \forall k,\ k \neq i, k \neq j$ onde $d^2(x_i, x_j)$ é o quadrado da distância euclideana entre x_i e x_j; ou seja, se nenhum outro ponto do conjunto pertence à bola de diâmetro $d(x_i, x_j)$ centrado no ponto médio do segmento entre x_i e x_j. Seja e um arco do grafo representativo dos dados e seja $c(e)$ o custo do arco e, dado pela distância entre os seus vértices extremos.

Sejam E_r^{ACM}, E_r^{GVR} e E_r^{GG} $(r = 1 \ldots, k)$, o conjunto de arcos que unem elementos da classe C_r, obtida por algum algoritmo de classificação aplicado ao conjunto de dados, no qual foi imposta uma estrutura de acordo com cada um dos três tipos de grafos. Sejam $\Delta_{ACM}(C_r) = max\{c(e_r^{ACM}) | e_r^{ACM} \in E_r^{ACM}\}$, $\Delta_{GVR}(C_r) = max\{c(e_r^{GVR}) | e_r^{GVR} \in E_r^{GVR}\}$ e $\Delta_{GG}(C_r) = max\{c(e_r^{GG}) | e_r^{GG} \in E_r^{GG}\}$. Estas medidas podem ser substituídas em (3.5) proporcionando mais três versões diferentes para o índice de *Dunn*. Quanto ao índice de *Davies-Bouldin*, podem obter-se três novas versões do índice se substituirmos

em (3.4) a dispersão de cada classe C_r $(r = 1, \ldots, k)$ por $\Delta_{ACM}(C_r)$, $\Delta_{GVR}(C_r)$ ou $\Delta_{GG}(C_r)$.

Da análise dos resultados de algumas experiências [183], concluiu-se que todas as generalizações do índice de *Dunn* mostraram ter um melhor desempenho do que o índice original, sendo o índice que usa o $\Delta_{GG}(C_r)$ o melhor. Os índices generalizados do índice de *Davies-Bouldin* mostraram-se mais eficientes na detecção de classes em forma de cadeia, e menos sensíveis na presença de pontos isolados.

3.5 Índice de *Krzanowski & Lai*

Sejam $\{x_{ij}\}$, $i = 1, \ldots, n$; $j = 1, \ldots, t$ os elementos a classificar. Suponhamos que o conjunto de dados foi classificado em $C_1, \ldots, C_k$ classes de $n_1, \ldots, n_k$ elementos, respectivamente. Em [134] é proposto o índice:

$$KL(k) = \left| \frac{DIFF(k)}{DIFF(k+1)} \right| \tag{3.8}$$

onde

$$DIFF(k) = (k-1)^{2/t} W_{k-1} - k^{2/t} W_k \tag{3.9}$$

e

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r \qquad \text{e} \qquad D_r = \sum_{i,i' \in C_r, i \neq i'} d_{ii'} \tag{3.10}$$

onde $d_{ii'} = \sum_{j=1}^{t} (x_{ij} - x_{i'j})^2$ é o quadrado da distância euclideana entre os elementos $x_i, x_i' \in C_r$. O valor de k indicativo do número de classes é aquele que maximiza o valor do índice, identificando assim uma partição na qual a soma das distâncias entre pares de elementos pertencentes à mesma classe, é a menor possível.

3.6 Estatística *Silhouette* de *Kaufman & Rousseeuw*

A estatística *Silhouette* de *Kaufman & Rousseeuw* proporciona um método gráfico de verificação da densidade e isolamento de classes, diferenciando entre elementos do centro e da fronteira da classe. A estatística é dada por ([199] e [131]):

$$s(x) = \frac{b(x) - a(x)}{max(a(x), b(x))} \tag{3.11}$$

onde $a(x)$ é a dissemelhança média do elemento x a todos os outros elementos pertencentes à mesma classe, $b(x)$ é a dissemelhança média do elemento x a todos os elementos pertencentes à classe que lhe está mais próxima, ou seja, à classe que minimiza esta média. Verifica-se $-1 \leq s(x) \leq 1$, sendo os valores negativos indicativos de que os elementos são similares aos membros de outras classes, e valores perto de 1 indicam que os elementos pertencem fortemente à classe na qual foram colocados, indicando que o elemento foi bem classificado. O valor indicativo do número de classes no conjunto de dados, é dado pelo valor que maximiza a média de $s(x)$ para todos os elementos do conjunto de dados.

3.7 *Gap Statistics*

Esta estatística pretende estimar o número de classes existente num conjunto de dados, comparando a dispersão intra-classe com o valor esperado sob o modelo nulo Unimodal. Ou seja [217],

$$Gap_n(k) = E_n^*(log(W_k)) - log(W_k) \tag{3.12}$$

onde W_k está definido em (3.10), E_n^* representa o valor esperado, para uma amostra de tamanho n, a partir da distribuição de referência. A estimativa do número de classes é o valor de k para o qual a diferença entre $log(W_k)$ e a curva de referência é maior, ou seja, é o valor que maximiza $Gap_n(k)$ após se considerar a distribuição da amostra. É conhecido que a inércia de uma classificação diminui à medida que o número de classes aumenta, podendo-se identificar o número de classes correcto aquele que corresponde a um "cotovelo" no gráfico. Esta estatística pretende formalizar esta heurística.

3.8 Estatística Modificada de Hubert

A *estatística modificada de Hubert* [112] é um índice de validação externo e mede o grau de semelhança entre o conjunto de dados representado pela sua matriz de dissemelhança, e a estrutura imposta por um algoritmo de classificação.

Seja $P = [p_{ij}]$ uma matriz de dissemelhança $n \times n$ e p_{ij} alguma medida de dissemelhança entre os elementos i e j. A matriz $Q = [q_{ij}]$ é uma matriz $n \times n$ definida por:

$$q_{ij} \quad = \quad \| \, r_{L(i)} - r_{L(j)} \, \|_2 \tag{3.13}$$

para $i, j = 1, 2, \ldots, n$; $L(i) = m$ se o elemento i pertence à classe m e $r_{L(i)}$ é o centro de gravidade da classe à qual o elemento i pertence.

A estatística de *Hubert* Γ, é definida a partir de P e Q e é dada por:

$$\Gamma(P, Q) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} p_{ij} q_{ij} \tag{3.14}$$

Na sua forma normalizada, Γ surge como um coeficiente de correlação simples entre as entradas de P e de Q:

$$\Gamma'(P, Q) = \frac{\frac{1}{M} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (p_{ij} - \bar{p})(q_{ij} - \bar{q})}{S_P S_Q} \tag{3.15}$$

onde $M = \dfrac{n(n-1)}{2}$ e:

$$\bar{p} = \frac{1}{M} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} p_{ij}, \qquad \bar{q} = \frac{1}{M} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} q_{ij}$$

$$S_P^2 = \frac{1}{M} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} p_{ij}^2 - \bar{p}^2, \qquad S_Q^2 = \frac{1}{M} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} q_{ij}^2 - \bar{q}^2$$

Para o índice normalizado tem-se $-1 \leq \Gamma' \leq 1$. O índice Γ' mede a relação linear entre as entradas de P e de Q.

Considere-se ainda, $m_1 = a + d$ (número de elementos que pertencem à mesma classe na primeira partição) e $m_2 = a + c$ (número de elementos que pertencem à mesma classe na segunda partição); mostra-se que [216]:

$$\Gamma' = \frac{Ma - m_1 m_2}{m_1 m_2 (M - m_1)(M - m_2)} \tag{3.16}$$

Experiências práticas [112] mostram que este índice tem tendência a aumentar com o número de classes. Sendo assim, não são os valores absolutos da estatística modificada que são examinados mas sim as mudanças verificadas nos valores como função do número de classes, identificando-se assim, o número de classes bem separadas, no "cotovelo" do gráfico representativo dos valores do índice contra o número de classes.

3.9 Índice de Lerman_Costa_Silva

Nesta secção é apresentado um índice que pretende determinar o número de classes existente num conjunto de dados, caracterizados por atributos numéricos e categóricos. Este critério resultou de uma linearização do índice de Lerman [140] e consequente extensão a atributos mistos ([141], [142] e [205]).

Consideremos um conjunto Ω de n elementos caracterizados por t atributos (ou variáveis) numéricos. Como resultado da aplicação de um método de classificação não hierárquica ao conjunto Ω, obtém-se uma partição em k classes. Designemos por C^ℓ uma dessas classes, n_ℓ o número de elementos da classe, g_ℓ o centro de gravidade da classe e I_ℓ o conjunto dos índices dos elementos da classe ℓ. O índice linear é dado por [142]:

$$\frac{1}{\sqrt{\frac{r \times s}{n^2}}} \sum_{1 \leq \ell \leq k} n_\ell \sum_{i \in I_\ell} \frac{d^2(x_i, g^\ell) - \mu_g}{\lambda_g} \tag{3.17}$$

onde

$$\mu_g = \frac{1}{n} \sum_{x \in \Omega} d^2(x, g) \qquad \text{e} \qquad \lambda_g^2 = \frac{1}{n} \sum_{x \in \Omega} \left\{ \left[d^2(x, g) \right]^2 \right\} - \mu_g^2 \tag{3.18}$$

são respectivamente a média e a variância das distâncias de todos os elementos do conjunto ao centro global, g.

Sejam ainda

$$r = \sum_{1 \leq \ell \leq k} n_l^2 \quad \text{e} \quad s = n^2 - r \tag{3.19}$$

O valor mínimo do índice é indicativo do número de classes existente no conjunto de dados.

O índice de Lerman_Costa_Silva LCS, consiste na adaptação do índice definido em (3.17) a conjuntos de dados com atributos mistos, podendo ser numéricos e/ou categóricos. Sejam $\Omega = \{x_1, \ldots, x_n\}$ o conjunto de dados e x_{ij} o valor do atributo j do objecto x_i. Considere-se que este valor é numérico se $1 \leq j \leq \ell$ e é categórico se $\ell + 1 \leq j \leq t$ $(t > \ell)$. A dissemelhança entre x_i e $x_{i'}$ pode ser calculada pela seguinte fórmula [111]:

$$d^2(x_i, x_{i'}) = \sum_{1 \leq j \leq \ell} (x_{ij} - x_{i'j})^2 + \gamma \sum_{\ell+1 \leq j \leq t} \delta(x_{ij}, x_{i'j}) \tag{3.20}$$

onde

$$\delta(x_{ij}, x_{i'j}) = \begin{cases} 1 & \text{se} \quad x_{ij} \neq x_{i'j} \\ 0 & \text{se} \quad x_{ij} = x_{i'j} \end{cases} \tag{3.21}$$

para $\ell + 1 \leq j \leq t$. O parâmetro de equilíbrio γ evita o favorecimento de um dos dois tipos de atributos, e é dado por 30% da média aritmética dos desvios padrão dos atributos numéricos.

O centro de gravidade do conjunto Ω adaptado a atributos mistos é dado por [111]:

$$g = (g_1, \ldots, g_h, \ldots, g_\ell, f_{\ell+1}, \ldots, f_{\ell+j}, \ldots, f_t) \tag{3.22}$$

onde g_h é a média da componente h em Ω, $1 \leq h \leq \ell$, e $f_{\ell+j}$ representa a moda da variável categórica $\ell + j$ (a categoria mais frequente em Ω), $1 \leq j \leq t - \ell$.

3.10 Aplicação do Índice de Lerman_Costa_Silva e Comparação com alguns Índices

Nesta secção pretende-se avaliar o desempenho do índice de Lerman_Costa_Silva LCS, aplicado a partições obtidas por um método de partição estendido a atributos mistos (*k-prototype*), comparando-o com alguns índices definidos em secções anteriores.

O algoritmo *k-prototype* [111] é uma extensão do algoritmo das *k-médias* suportando atributos mistos (numéricos e/ou categóricos). Neste algoritmo, a dissemelhança entre dois elementos é dada por (3.20). Os centros de gravidade são determinados pela média dos atributos numéricos e pela moda dos atributos categóricos (3.22).

3.10.1 Geração dos Conjuntos de Dados

Os conjuntos de dados usados nesta secção foram gerados como se segue. O gerador de dados produz classes de formas esférica e elíptica com densidades de pontos diferentes no interior de cada classe. Cada classe é composta por três círculos ou elipses com o mesmo centro mas com diferentes raios ou parâmetros elípticos. Em cada arco de círculo ou de elipse os pontos são atribuídos com diferentes probabilidades, podendo algumas delas ser zero. Para a geração de uma classe esférica deverão ser fornecidos os seguintes parâmetros, representados na figura 3.1:

$$< p >< a >< b >< r_1 >< r_2 >< r_3 >< p_1 >< p_2 >< p_3 >$$

onde p é a percentagem de elementos que deverão pertencer a esta classe, (a, b) é o centro das três circunferências, de raios r_1, r_2 e r_3. O valor p_1 é a percentagem dos p elementos que pertencem ao círculo (A_1), p_2 é a percentagem dos p elementos que pertencem à coroa circular central (A_2) e p_3 é a percentagem dos p elementos que pertencem à coroa circular mais exterior (A_3).

Para a geração de classes elípticas deverão ser fornecidos os seguintes parâmetros, representados na figura 3.2:

$$< p >< a >< b >< a_1 >< b_1 >< a_2 >< b_2 >< a_3 >< b_3 >< p_1 >< p_2 >< p_3 >$$

onde p é a percentagem de elementos que deverão pertencer a esta classe, (a, b) é o centro das três elipses, cujos parâmetros são (a_1, b_1), (a_2, b_2) e (a_3, b_3). O valor p_1

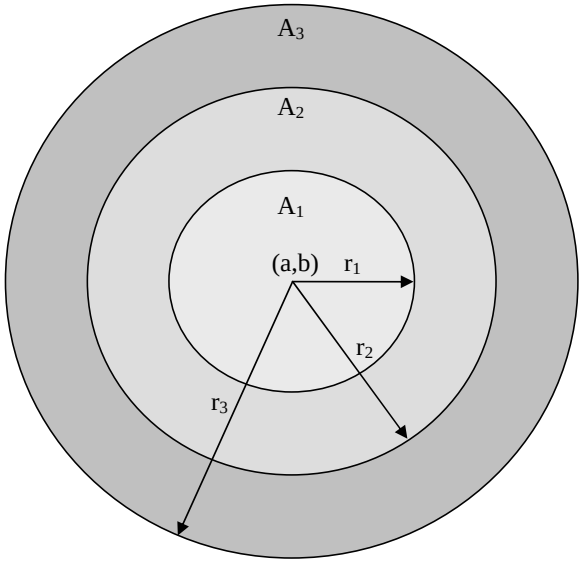

Figura 3.1: Esquema de uma classe esférica obtida pelo gerador de classes artificiais.

é a percentagem dos p elementos que pertencem à elipse mais interior (E_1), p_2 é a percentagem dos p elementos que pertencem à coroa elíptica central (E_2) e p_3 é a percentagem dos p elementos que pertencem à coroa elíptica mais exterior(E_3).

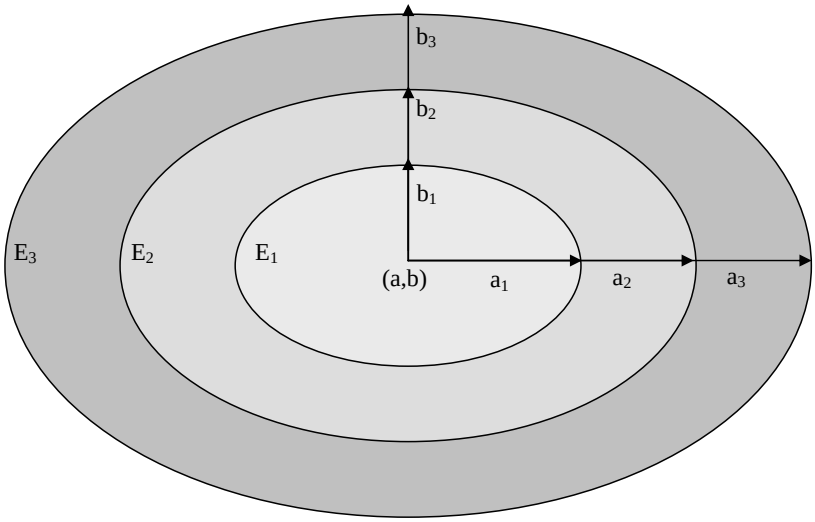

Figura 3.2: Esquema de uma classe elíptica obtida pelo gerador de classes artificiais.

O gerador de números aleatórios entre dois números inteiros fixados foi obtido na página da Internet: www.planet-source-code.com.

3.10.2 Descrição dos Conjuntos de Dados

O primeiro conjunto de dados está representado na figura 3.3, é composto por 20000 elementos caracterizados por dois atributos numéricos num espaço euclideano, distribuídos por cinco classes. A tabela 3.1 indica os parâmetros do gerador de dados para as cinco classes do conjunto. A figura 3.4 representa a partição em cinco classes identificada pelo método *k-prototype*.

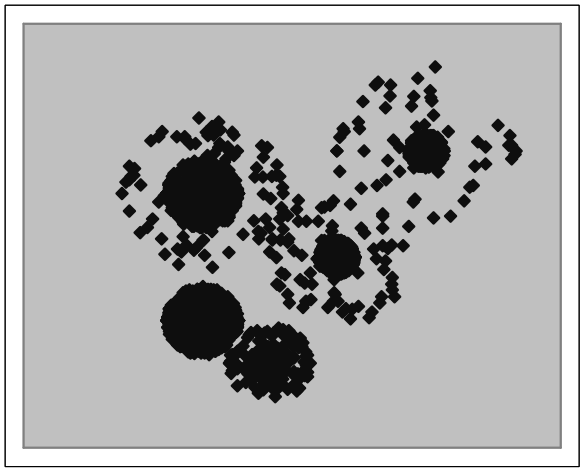

Figura 3.3: Conjunto com 20000 elementos caracterizados por dois atributos numéricos.

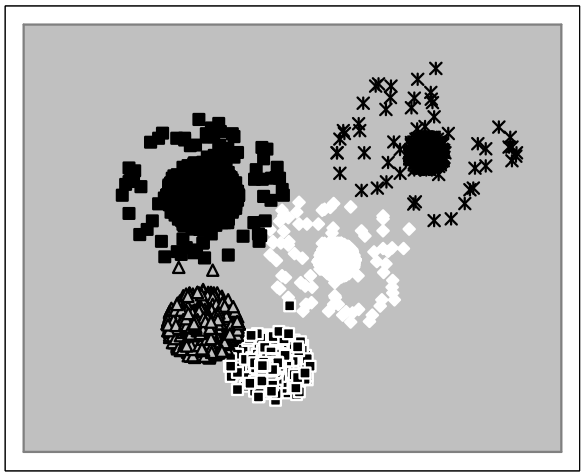

Figura 3.4: Distribuição em cinco classes identificadas pelo método *k-prototype*.

O segundo conjunto de dados deriva do primeiro, aumentando a dimensão com quatro atributos categóricos com cinco categorias cada, seguindo a mesma distribuição

Tabela 3.1: Parâmetros do gerador de dados para as cinco classes do conjunto de dados representado na figura 3.3

p	a	b	r_1	r_2	r_3	p_1	p_2	p_3
30	100	100	25	25	25	100	0	0
25	100	180	20	35	50	95	3	2
10	140	65	15	20	25	80	15	5
20	190	150	10	20	50	95	3	2
15	225	225	10	20	50	95	3	2

de probabilidades em elementos das mesmas classes, e distribuições diferentes entre elementos de classes diferentes. Pretende-se deste modo, que a introdução dos quatro novos atributos, não modifique a distribuição dos elementos nas classes. A distribuição de probabilidades das diferentes categorias dos atributos categóricos, está representada na tabela 3.2.

3.10.3 Resultados

O índice de LCS var ser comparado com os índices SD, DBI, DI e KL, definidos em secções anteriores. No caso do índice DI optou-se pela média das dissemelhanças entre pares de elementos de diferentes classes para definir a dissemelhança entre classes. O diâmetro de uma classe é dado pela dissemelhança máxima entre pares de elementos da classe.

O algoritmo *k-prototype* foi executado para $k = 2$ até $k = 15$, sendo para cada uma das partições calculados os índices indicados.

A figura 3.5 representa os valores do índice LCS, obtidos para cada uma das partições do primeiro conjunto de dados caracterizado por dois atributos numéricos. Como se pode ver pelo gráfico, o valor mínimo corresponde a $k = 5$ cuja partição está representada na figura 3.4.

A tabela 3.3 representa os resultados obtidos para todos os índices seleccionados. No caso dos índices SD, DBI e LCS, o valor mínimo dos índices é indicativo do número de classes, enquanto que para os índices DI e KL, o número de classes a considerar

Tabela 3.2: Distribuição de probabilidades das categorias dos quatro atributos categóricos, para cada uma das cinco classes.

Atributo 1				
'a'	'b'	'c'	'd'	'e'
0.072	0.116	0.115	0.642	0.055
0.070	0.570	0.186	0.112	0.062
0.075	0.190	0.119	0.122	0.494
0.080	0.188	0.566	0.132	0.034
0.071	0.170	0.581	0.124	0.054

Atributo 2				
'f'	'g'	'h'	'i'	'j'
0.066	0.109	0.668	0.078	0.079
0.084	0.172	0.142	0.503	0.099
0.070	0.101	0.596	0.131	0.102
0.091	0.114	0.568	0.145	0.082
0.060	0.698	0.040	0.127	0.075

Atributo 3				
'k'	'l'	'm'	'n'	'o'
0.068	0.528	0.113	0.204	0.087
0.074	0.470	0.162	0.207	0.087
0.061	0.096	0.557	0.193	0.093
0.076	0.139	0.504	0.205	0.076
0.064	0.135	0.535	0.186	0.080

Atributo 4				
'p'	'q'	'r'	's'	't'
0.077	0.160	0.217	0.495	0.051
0.065	0.553	0.265	0.058	0.059
0.071	0.186	0.172	0.082	0.489
0.063	0.155	0.710	0.048	0.024
0.082	0.216	0.165	0.459	0.078

Tabela 3.3: Resultados da aplicação dos índices às partições obtidas pelo método *k-prototype*, para o conjunto de atributos numéricos.

	SD	DBI	DI	KL	LCS
$k = 2$	0.392	0.795	0.446	2.030	−10184.82
$k = 3$	0.311	0.699	0.428	0.843	−11402.02
$k = 4$	0.176	0.423	0.298	0.940	−11216.69
$k = 5$	**0.089**	**0.162**	**0.454**	**5.637**	**−11610.54**
$k = 6$	0.306	0.498	0.095	0.918	−10991.47
$k = 7$	1.503	0.557	0.092	0.986	−10817.02
$k = 8$	2.559	0.620	0.079	1.118	−10678.32
$k = 9$	0.500	0.757	0.079	1.571	−10528.74
$k = 10$	0.404	0.888	0.078	0.755	−9918.87
$k = 11$	0.480	0.870	0.077	0.951	−9719.78
$k = 12$	2.689	0.873	0.069	2.870	−9608.35
$k = 13$	0.619	0.958	0.069	2.590	−7096.19
$k = 14$	2.228	0.930	0.069	0.138	−6195.30
$k = 15$	0.680	0.948	0.069	0.419	−6355.76

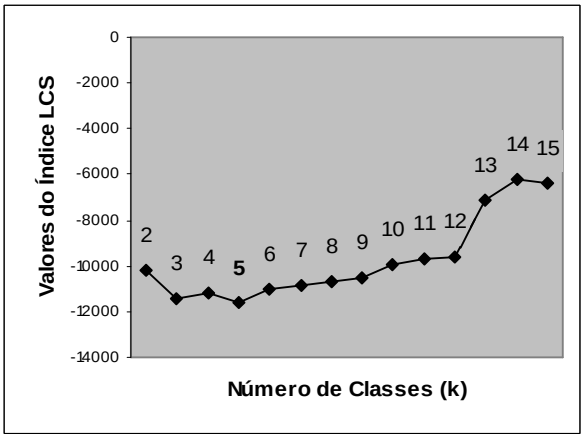

Figura 3.5: Gráfico representativo dos valores do índice LCS para o conjunto de dados com atributos numéricos, representado na figura 3.3.

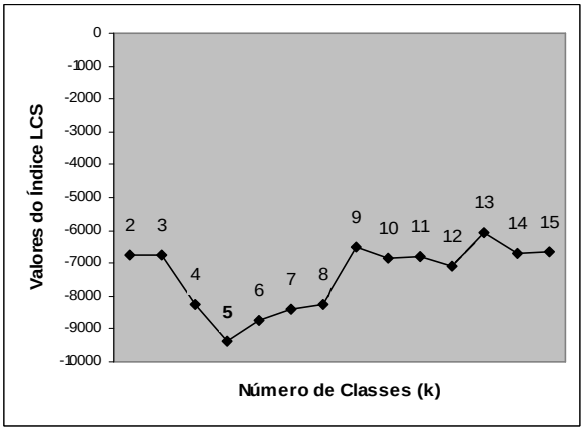

Figura 3.6: Gráfico representativo dos valores do índice LCS para o conjunto de dados com atributos mistos.

corresponde ao valor máximo apresentado por estes índices. Como se pode ver pela tabela, todos os índices foram eficazes na identificação da partição em cinco classes como aquela que melhor representa o conjunto de dados. No entanto, quanto aos tempos de execução os resultados são diferentes. A tabela 3.4 representa o tempo de execução dos índices para as 14 partições analisadas, sendo a execução feita num PC *Pentium 4* de 3GHz a 512 Mbytes de RAM. Como se pode ver pela referida tabela, o índice LCS apresenta o terceiro melhor tempo, tendo dois índices com pior tempo, os índices DI e KL e dois com melhor tempo, os índices SD e DBI.

No caso do segundo conjunto com dados de atributos mistos, o gráfico representativo dos valores do índice LCS está representado na figura 3.6, identificando também a

Tabela 3.4: Tempos de execução dos índices para as 14 partições analisadas.

	Tempo (seg.)
SD	0.34
DBI	0.52
DI	233.05
KL	70.14
LCS	30.66

partição em cinco classes.

Para analisar o segundo conjunto de dados com atributos mistos, foi feita uma extensão de cada um dos índices para atributos mistos usando a mesma abordagem, quanto à função dissemelhança entre elementos e centros de gravidade, que a usada para o índice LCS, definindo os índices SDM, DBIM, DIM e KLM.

Obtiveram-se os resultados representados na tabela 3.5. Como se pode ver pela tabela só os índices KLM e LCS identificaram a partição em cinco classes. Tendo em conta que para o conjunto de atributos numéricos, o índice KLM foi mais lento que o índice LCS, os resultados mostram que relativamente a atributos mistos, o índice LCS é mais eficiente que os índices considerados, estendidos a atributos mistos.

3.11 Determinação do número de classes nos métodos hierárquicos: Regras de Corte

Nos métodos hierárquicos, os procedimentos que procuram o valor mais apropriado de k, são em geral referidos como *regras de corte*. Usando estes métodos podemos seguir dois caminhos. Por um lado, pode obter-se uma classificação hierárquica completa, seguindo-se a aplicação de um indicador do nível de "corte" da estrutura em árvore, de modo a obter-se o número de classes [8]. Ou por outro lado, pode existir um parâmetro externo que, num determinado momento da execução do método, faça parar a reunião ou divisão de classes, parando assim o processo de classificação. Uma terceira abordagem consiste em inserir no próprio processo de classificação uma condição a satisfazer para que as classes sejam reunidas ([74] e [73]) ou separadas [81]. Em [162]

Tabela 3.5: Resultados da aplicação dos índices estendidos a atributos mistos, às partições obtidas pelo método *k-prototype* para o conjunto de atributos mistos.

	SDM	DBIM	DIM	KLM	LCS
$k = 2$	**0.0046**	**2.617**	**0.171**	2.464	−6727.16
$k = 3$	0.0046	2.829	0.160	2.169	−6745.33
$k = 4$	0.0046	2.827	0.157	0.369	−8240.39
$k = 5$	0.0048	2.873	0.153	**2.760**	**−9348.47**
$k = 6$	0.0047	2.806	0.137	2.470	−8725.75
$k = 7$	0.0050	2.848	0.127	0.409	−8388.48
$k = 8$	0.0055	3.084	0.122	2.445	−8243.49
$k = 9$	0.0054	2.941	0.120	0.492	−6491.12
$k = 10$	0.0055	2.948	0.121	2.062	−6864.70
$k = 11$	0.0058	3.132	0.112	0.468	−6780.59
$k = 12$	0.0059	2.984	0.108	1.682	−7081.56
$k = 13$	0.0062	2.924	0.107	0.863	−6082.39
$k = 14$	0.0074	2.938	0.099	1.951	−6678.41
$k = 15$	0.0084	3.156	0.110	0.419	−6658.90

é apresentado um trabalho de comparação entre 30 regras de corte das quais as cinco melhores classificadas estão representadas na tabela 3.6.

3.11.1 Índice de *Calinski & Harabasz*

O índice é dado por [39]:

$$CH(k) = \frac{B(k)/(k-1)}{W(k)/(n-k)} \tag{3.23}$$

onde

$$W(k) = \sum_{l=1}^{k} \sum_{x \in C_r} d^2(x, g_r) \quad \text{e} \quad B(k) = \sum_{i,j=1, i \neq j} dist^2(C_i, C_j) \tag{3.24}$$

sendo g_r o centro de gravidade da classe r e $dist(C_i, C_j)$ é a dissemelhança entre classes. Escolhe-se o valor de k que maximiza o valor do índice *Calinski & Harabasz*,

Tabela 3.6: Regras de corte consideradas as melhores em [162].

Nome	Fórmula	Referência
Índice de Calinski & Harabasz	$\frac{\frac{B(k)}{k-1}}{\frac{W(k)}{n-k}}$	[39]
Índice de Duda & Hart	$\frac{Je(2)}{Je(1)}$	[54]
Índice C-Index	$\frac{d_w - min(d_w)}{max(d_w) - min(d_w)}$	[117]
Índice Gamma	$\frac{s(+) - s(-)}{s(+) + s(-)}$	[10]
Índice de Beale	$F = \frac{\frac{W_1 - W_2}{W_2}}{\frac{k-1}{k-2}2^{\frac{2}{t}} - 1}$	[91]

identificando-se assim uma partição em que soma das dissemelhanças entre os elementos e o centro da classe à qual pertencem é menor que a soma das distâncias entre classes, ou seja, identifica classes compactas e isoladas.

Note-se que pela definição deste índice, este pode ser usado num contexto não hierárquico, tal como os índices referidos em secções anteriores, aplicado a diferentes partições do mesmo conjunto de dados.

3.11.2 Índice de *Duda & Hart*

Este critério é aplicado a algoritmos que seguem uma abordagem ascendente, e pretende decidir sobre a reunião ou não de duas classes no processo de classificação. O índice é baseado na razão entre a soma dos quadrados das dissemelhanças intra-classes das duas classes candidatas a serem reunidas ($J(2)$), e a soma dos quadrados das dissemelhanças da classe resultante da reunião ($J(1)$). A hipótese de uma classe é rejeitada, ou seja, as duas classes candidatas não se reunem, se a razão for menor que um dado parâmetro.

3.11.3 Índice *C-Index*

Este índice é dado por [117]: $\frac{d_w - min(d_w)}{max(d_w) - min(d_w)}$, onde d_w é a soma das dissemelhanças intra-classes. O valor mínimo ao longo dos níveis da hierarquia é usado para indicar o número de classes óptimo. Em [160] prova-se que este índice é muito eficaz na identificação do número de classes existente no conjunto de dados.

3.11.4 Índice *Gamma*

Este índice representa uma adaptação da estatística γ de *Goodman and Kruskal* ([87]) para ser usada no contexto da análise classificatória [10]. Este índice já foi referido na Secção 2.4, é calculado para cada nível da hierarquia, sendo o nível de corte dado para valores máximos do índice.

3.11.5 Índice de *Beale*

Este índice aplica-se a métodos hierárquicos descendentes, e usa um *F-ratio* para testar a hipótese da existência de k_2 contra k_1 classes no conjunto de dados ($k_2 > k_1$). O *F-ratio* avalia o aumento da média dos quadrado dos desvios dos valores dos elementos das classes em relação aos centróides, quando se vai de k_2 para k_1 classes, contra a média dos quadrados dos respectivos desvios da partição com k_2 classes. As tabelas F podem ser usadas com $t(k_2 - k_1)$ e $t(n - k_2)$ graus de liberdade, onde t é o número de atributos e n é o número de elementos no conjunto de dados. Em cada nível da hierarquia é feito um teste de duas contra uma classes, continuando a classificação até que a hipótese de uma classe seja rejeitada. A estatística de teste para decidir se uma classe deve ser dividida é dado por:

$$F = \frac{\frac{W_1 - W_2}{W_2}}{\frac{m-1}{m-2}2^{\frac{2}{t}} - 1} \tag{3.25}$$

(onde W_1, W_2, m e t são definidos como para o índice anterior) com uma distribuição $F_{t,(k-2)t}$, sob a hipótese nula, admitindo multinormalidade dos atributos. Rejeita-se a

hipótese de uma classe única para valores significativamente grandes de F.

3.11.6 Índices de Validação Interna

Em [203] são apresentados alguns índices de validação interna que podem ser aplicados a cada passo de um algoritmo de classificação hierárquica ascendente, com o objectivo de em conjunto, orientarem na determinação do número de classes existente no conjunto de dados (ver [102] para uma aplicação). Os índices são conhecidos como: *Root-mean square standard deviation* de uma nova classe, mede a homogeneidade das classes formadas em cada passo que deverá ser menor do que no passo anterior; *Semi-partial R-squared*, mede a perda de homogeneidade devido à reunião de duas classes, se o valor do índice for zero então a nova classe é obtida pela reunião de duas classes perfeitamente homogéneas; *R-squared*, mede do grau de diferença existente entre grupos, um valor zero indica que não existe diferença entre grupos enquanto que o valor 1 indica diferenças significativas entre grupos; Distância entre Classes, mede a distância entre duas classes que são reunidas num determinado passo.

3.12 Conclusão

A determinação do número de classes de um conjunto de dados numa análise não supervisionada, não é um problema de solução fácil devido à grande diversidade dos conjuntos de dados, nomeadamente quanto à forma das classes ou ao tipo de atributos. Esta questão é abordada de modo diferente de acordo com o tipo de método de análise classificatória: partição ou hierárquico. No caso hierárquico, têm sido desenvolvidos índices que param a reunião ou a separação de classes no processo hierárquico, seguindo uma abordagem ascendente ou descendente respectivamente. Relativamente aos métodos de partição, as soluções que têm surgido na literatura passam por desenvolver índices de validação, que permitem medir a qualidade de uma partição. Fazendo variar o valor de k entre limites pré-definidos, aqueles índices pretendem identificar a partição que melhor reflecte a estrutura existente no conjunto de dados. Neste contexto, foi apresentado o índice de LCS que associado ao método de partição *k-prototype*, pretende identificar o número de classes existente no conjunto de dados caracterizados por atributos mistos. O método proposto mostrou ser tão eficiente como alguns índices existentes na literatura, no caso de atributos numéricos,

mostrando-se superior quando aplicado a atributos mistos.

Capítulo 4

Um Método de Partição Validado por um Índice de Classificabilidade

Aplicando a teoria de grafos à análise classificatória, é proposto um método de classificação não hierárquica validado por um índice de classificabilidade. O método não requer qualquer pressuposto sobre a distribuição dos dados. Os vértices do grafo definido sobre o conjunto de dados, correspondem aos elementos a classificar; enquanto que os arcos unem dois vértices se os elementos correspondentes do conjunto distarem mais do que o valor de um parâmetro de controlo. O método é baseado na coloração de grafos e usa um algoritmo sequencial ajustado. O número e a composição das classes que melhor representam a estrutura existente no conjunto de dados, são aqueles que optimizam o valor do índice de classificabilidade. A ideia chave deste índice é a de que existem k classes compactas e bem separadas, se for possível definir um grafo k-*partite* completo nos dados, após a classificação. O índice de classificabilidade também permite identificar conjuntos de dados sem uma estrutura em classes.

A eficácia do método é constatada em alguns conjuntos de dados artificiais, evidenciando a sua robustez no que diz respeito a conjuntos de dados com pontos isolados, com classes de formato não esférico e sem uma estrutura em classes. A estabilidade do método é avaliada, reclassificando um conjunto de dados modificado, e avaliando a extensão das diferenças em relação à classificação original.

Os algoritmos sequenciais de coloração dos vértices de um grafo não são óptimos, no entanto, se se escolher uma ordem adequada dos vértices a colorir, pode obter-se uma coloração com o menor número de cores possível. Tendo este facto em conta, foram

testados alguns algoritmos de coloração sequenciais, com o objectivo de seleccionar aquele que melhor responde ao requisito de identificar classes homogéneas no conjunto de dados.

Finalmente, a complexidade de execução do método é avaliada, analisando o tempo de execução médio e o tempo de execução no pior caso.

4.1 Noções de Teoria de Grafos

Seja $V(G) = \{v_1, v_2, \ldots, v_n\}$ um conjunto de n *vértices* e

$$E(G) = \{e_1, ..., e_m\} = \{e_\ell = \{v_i, v_j\} | v_i, v_j \in V, v_i \neq v_j, \ell \in \{1, \ldots, m\}\} \tag{4.1}$$

o conjunto de m *arcos* do grafo $G(V, E)$.

Sejam $v_i, v_j \in V$ dois vértices diferentes do grafo G; se $\exists \ell \in \{1, \ldots, m\} : e_\ell = \{v_i, v_j\}$ então os vértices v_i e v_j dizem-se *adjacentes* enquanto e_ℓ e v_i ou e_ℓ e v_j dizem-se *arco e vértice incidentes*.

Para qualquer grafo G, verifica-se $m \leq \left(\begin{array}{c} n \\ 2 \end{array} \right)$ e $E \subseteq \{\{v_i, v_j\}, \forall i, j \in \{1, 2, \ldots, n\}, i \neq j\}$. No caso de $m = \left(\begin{array}{c} n \\ 2 \end{array} \right)$ e $E \equiv \{\{v_i, v_j\}, \forall i, j \in \{1, 2, \ldots, n\}, i \neq j\}$, todos os vértices do grafo são adjacentes e este denomina-se *grafo completo*.

O número de arcos de G, incidentes com um vértice v designa-se *grau do vértice* v e representa-se por $d_G(v)$, ou seja

$$d_G(v) = Card\{u \in V : \{u, v\} \in E(G)\},\ v \in V(G) \tag{4.2}$$

Os graus dos vértices de um grafo relacionam-se com o número de arcos da seguinte forma:

$$\sum_{v \in V(G)} d_G(v) = 2m \tag{4.3}$$

Define-se em seguida *subgrafo*, *grafo induzido* e *grafo parcial*, de um grafo $G(V, E)$; assim como *vizinhança* de um vértice e de um conjunto de vértices.

Definição 4.1 *Seja $G(V, E)$ um grafo com n vértices e m arcos. Seja $V' \subset V$ um subconjunto do conjunto dos vértices de G, e seja $E' \subset E$ um subconjunto do conjunto dos arcos de G unindo vértices em V', o grafo $G'(V', E')$ designa-se por* Subgrafo *de G.*

Definição 4.2 *Seja $G(V, E)$ um grafo com n vértices e m arcos. Seja $V' \subset V$ um subconjunto do conjunto dos vértices de G, o grafo com V' como conjunto de vértices e cujos arcos são aqueles que unem os vértices de V' em G, designa-se* Grafo Induzido por V'.

Definição 4.3 *Seja $G(V, E)$ um grafo com n vértices e m arcos. Seja $E' \subset E$ um subconjunto do conjunto dos arcos de G, o grafo $G'(V, E')$ designa-se por* Grafo Parcial *de G.*

Definição 4.4 *Seja $v \in V(G)$ um vértice do grafo $G(V, E)$. A* Vizinhança de v, *$Viz(v)$, é o conjunto dos vértices de G adjacentes com v, ou seja,*

$$Viz(v) := \{u \in V(G) : \{u, v\} \in E(G)\} \tag{4.4}$$

Definição 4.5 *Seja $W \subset V(G)$ um subconjunto de vértices do grafo $G(V, E)$. A* Vizinhança de W, *$Viz(W)$, é o conjunto dos vértices de G adjacentes com algum vértice de W, ou seja,*

$$Viz(W) := \{u \in V(G) : \exists v \in W, \{u, v\} \in E(G)\} \tag{4.5}$$

4.1.1 Classes de Cor

Uma das propriedades mais importantes e talvez um dos problemas mais estudados no contexto da teoria de grafos, é o problema da *coloração de grafos* ([34], [202], [150], [46], [123], [59], [85], [41], [126], [83] e [63]). A coloração de grafos envolve a atribuição de cores aos vértices de um grafo de tal modo que a dois vértices adjacentes sejam atribuídas cores diferentes, com o objectivo de minimizar o número de cores usado. Dada uma coloração de um grafo G, o conjunto que consiste em todos os vértices coloridos com uma mesma cor é referido como *classe de cor*. Ou seja, é sempre possível encontrar uma partição do conjunto de vértices V, de modo que qualquer par

de vértices pertencentes à mesma parte, ou classe de cor, é não adjacente. No caso limite de um grafo completo de n vértices, são necessárias n cores para o colorir, uma vez que qualquer par de vértices é adjacente. O número mínimo de cores k ($k \leq n$), necessárias e suficientes para colorir um grafo G é denominado *número cromático* e representa-se por $\chi(G)$.

Definamos de modo mais formal uma *coloração de um grafo G*.

Definição 4.6 *([123], [94]) Dado um grafo $G(V, E)$ com V como conjunto dos vértices e E como conjunto dos arcos, a função*

$$\omega : V \to \{1, \ldots, k\}$$

é uma k-coloração *se $\omega(u) \neq \omega(v)$ para todos os arcos $\{u, v\} \in E$. Ou seja, define-se uma* k-coloração *se for possível colorir todos os vértices com uma das k cores, de modo que cada arco une dois vértices de cores diferentes.*

Considere-se que o grafo G foi colorido com k cores, sejam C_i com $i = 1, \ldots, k$ as k classes de cor, assim

$$C_i = \{v \in V : \omega(v) = i\} \tag{4.6}$$

As classes de cor verificam as seguinte propriedades:

1. $C_i \cap C_j = \emptyset,\ \forall i, j \in \{1, \ldots, k\}$ e $i \neq j$
2. $\bigcup_{i=1}^{k} C_i = V(G)$
3. $\forall i \in \{1, \ldots, k\} \forall v \in C_i,\ Viz(v) \subseteq V \setminus C_i$
4. Se $u, v \in C_i \Rightarrow \{u, v\} \notin E(G)$

Se o algoritmo de coloração dos vértices tiver uma execução óptima, verifica-se $Viz(v) \equiv V \setminus C_i$, ou seja, cada vértice é não adjacente aos vértices pertencentes à mesma classe de cor, sendo adjacente a todos os vértices fora da classe. No entanto, o facto de não serem conhecidos algoritmos de coloração óptima de execução polinomial, condiciona a identificação de tais colorações, permitindo que vértices não adjacentes sejam colocados em classes diferentes. Pelas mesmas razões, notar que a condição (4) é não suficiente.

4.1.2 Grafos *k-Partite*

Um grafo *k-partite* é um grafo cujo conjunto de vértices pode ser dividido em k conjuntos de vértices não adjacentes; sendo *k-partite completo* se todos os vértices pertencentes a conjuntos diferentes forem adjacentes. Deste modo, um grafo *k-partite* é constituído por k classes de cor.

Definição 4.7 *Se $P = \{C_1, \ldots, C_k\}$ for uma partição do conjunto V dos vértices de um grafo $G(V, E)$, com $Card\{C_i\} = n_i$ para $i = 1, \ldots, k$, o grafo é k-partite* completo *e representa-se por $K_{n_1,n_2,\ldots,n_k}$ se*

$$\forall i \in \{1, \ldots, k\} \, \forall v \in C_i, \ \{u, v\} \in E(G), \ \forall u \in C_j \ \ \forall j \neq i \tag{4.7}$$

Ou seja, todos os vértices pertencentes a conjuntos diferentes são adjacentes.

4.1.3 Multigrafos

Os *multigrafos* são grafos de arcos múltiplos, ou seja, entre dois vértices pode existir mais do que um arco. Estes grafos fazem sentido se associarmos significados diferentes a cada um dos arcos. Por exemplo, neste trabalho aplicam-se os multigrafos para analisar dados com atributos diferenciados.

4.1.4 Conjuntos Independentes Maximais

Comecemos por definir conjuntos independentes no contexto da teoria de grafos.

Definição 4.8 *Seja $G(V, E)$ um grafo, um subconjunto $I \subset V$ diz-se* Conjunto Independente *se*

$$\forall u, v \in I, \ \ \{u, v\} \notin E(G)$$

Ou seja, um conjunto independente é composto por vértices não adjacentes.

Pelo já exposto, as classes de cor são conjuntos independentes. Se se obtiver uma coloração óptima do conjunto de dados, as classes de cor são conjuntos independentes maximais, no sentido em que cada classe tem o maior número de vértices possível.

Definição 4.9 *Seja $G(V, E)$ um grafo, um subconjunto $I \subset V$, diz-se* Conjunto Independente Maximal *se não existir nenhum vértice $v \in V \setminus I$ tal que $I \cup \{v\}$ é conjunto independente, ou seja,*

$$\forall v \in V \setminus I, \ \ \exists u \in I : \ \{u, v\} \in E$$

4.2 Método de Classificação Baseado em Classes de Cor

Nesta secção, é apresentado um algoritmo de classificação não hierárquica baseado em classes de cor. O algoritmo de coloração usado é sequencial (*greedy*), e é seguido por um processo de ajuste de modo a reduzir o número de cores encontradas, orientado pela procura de classes homogéneas.

4.2.1 Definição do Grafo Representativo dos Dados

Seja $\Omega = \{x_1, x_2, \ldots, x_n\}$ o conjunto de n elementos a classificar. Seja ϕ uma função bijectiva de Ω em V:

$$\begin{array}{rcccc} \phi & : & \Omega & \longrightarrow & V \\ & & x & \hookrightarrow & v \end{array} \tag{4.8}$$

que a cada elemento do conjunto de dados faz corresponder um vértice de um grafo $G(V, E)$, definido sobre o conjunto de dados. Defina-se uma medida de dissemelhança d sobre Ω:

$$\begin{array}{rcccc} d & : & \Omega \times \Omega & \longrightarrow & \mathbb{R}_0^+ \\ & & (x_i, x_j) & \hookrightarrow & d(x_i, x_j) \end{array} \tag{4.9}$$

Representemos por *dist* a dissemelhança entre vértices do grafo correspondentes aos elementos do conjunto. Se u é tal que $\phi(x_i) = u$ para algum $x_i \in \Omega$ e v é tal que $\phi(x_j) = v$ para algum x_j em Ω, então $dist(u, v) := d(x_i, x_j)$.

O conjunto dos arcos é definido por

$$E = \{\{u, v\} : u, v \in V; dist(u, v) \geq \alpha\} \tag{4.10}$$

sendo α um parâmetro de controlo. Ou seja, existe um arco a unir dois vértices diferentes do grafo se a sua dissemelhança for superior ou igual ao valor do parâmetro de controlo.

Definição 4.10 *Seja $G(V, E)$ o grafo definido no conjunto de dados $\Omega = \{x_1, x_2, \ldots, x_n\}$, tal que $\phi(x_i) = v_i,\ \forall x_i \in \Omega$ e $E = \{\{v_i, v_j\} : v_i, v_j \in V : dist(v_i, v_j) \geq \alpha\}$, então para $v_i, v_j \in V$ verifica-se*

$$\{v_i, v_j\} \in E \Longleftrightarrow dist(v_i, v_j) \geq \alpha \tag{4.11}$$

Ou seja, os vértices adjacentes são aqueles cujos elementos correspondentes do conjunto de dados estão a uma dissemelhança superior ou igual ao valor de um parâmetro α.

4.2.2 Algoritmo de Coloração Sequencial

Na literatura muitos algoritmos têm sido propostos para resolver o problema da coloração de grafos ([34], [46], [123], [41] e [83]), no entanto, nenhum deles pode garantir uma atribuição de cores óptima para os vértices do grafo. O problema de encontrar uma coloração com o menor número possível de cores é um problema *NP-Completo* [82], ou seja, não se conhece nenhum algoritmo de coloração óptima de execução polinomial. A execução exponencial torna a sua aplicação impossível a conjuntos de dados com mais de algumas centenas de elementos. Assim, deverá optar-se por um algoritmo de coloração que obtenha uma atribuição de cores em maior número que o número cromático. Um dos algoritmos não óptimos mais simples e mais usados em teoria de grafos para obter uma coloração de um grafo, designa-se por *Algoritmo de Coloração Sequencial* (ACS) ("greedy algorithm"). Este algoritmo começa por atribuir ao primeiro vértice a primeira cor. Os restantes vértices são coloridos, de modo sequencial, com a primeira cor permitida, ou seja, a primeira das cores ainda não atribuída a nenhum dos seus vértices adjacentes.

O algoritmo segue então os seguintes passos:

1. Seja $v_1 \in V(G)$
2. Seja $\omega(v_1) = 1$
3. Seja C_i a classe de cor i e suponhamos já coloridos j vértices em k cores
4. Se $\exists i : C_i \cap Viz(v_{j+1}) = \emptyset$
   então $\omega(v_{j+1}) = i$
   senão $\omega(v_{j+1}) = k + 1$

onde $Viz(v_{j+1})$ é a vizinhança de v_{j+1}.

A eficácia deste algoritmo de coloração sequencial depende da ordem de entrada dos dados. Se π for uma permutação do conjunto de vértices a colorir, então $\chi_s(G,\pi)$ representa o número de cores obtido pelo algoritmo de coloração para a permutação π. O número de classes de cor obtidas com a aplicação deste algoritmo é maior ou igual ao número cromático, ou seja,

$$\chi_s(G,\pi) \geq \chi(G) \tag{4.12}$$

No entanto, sabe-se [45] que existe uma ordem dos dados para o qual a coloração obtida pelo algoritmo sequencial é óptima, ou seja, $\exists\pi'$ tal que:

$$\chi_s(G,\pi') = \chi(G) \tag{4.13}$$

Sabe-se ainda [59] que se $\Delta = max_{\{v\in V\}}d_G(v)$ então o algoritmo de coloração sequencial usa no máximo $\Delta+1$ cores, usando em média o dobro do número de cores mínimo [29].

4.2.3 Método de Partição baseado na Coloração de Grafos

Devido à execução não óptima do algoritmo sequencial escolhido, quase todas as atribuições de cores requerem um número de cores superior ao necessário e suficiente. Como o objectivo é estabelecer uma correspondência entre as classes de cor e as classes homogéneas do conjunto de dados, um número excessivo de cores pode fragmentar as classes homogéneas. Este facto é pertinente uma vez que em [105], se prova que as classes identificadas por uma coloração óptima, correspondem a uma partição com diâmetro mínimo. No entanto, tendo como critério de classificação o diâmetro, o método favorece a identificação de classes esféricas em detrimento das alongadas. Como um dos objectivos do método proposto é conseguir identificar também classes de forma alongada, o algoritmo de coloração exponencial óptimo não responde aos requisitos pretendidos. Sendo assim, a redução do número de classes deverá ser feita baseada na densidade das classes e não no seu diâmetro.

O *Método de Partição baseado na Coloração de Grafos* (MPCG) consiste num ajuste da coloração obtida pelo *Algoritmo de Coloração Sequencial* (ACS), reduzindo o número de cores, com o objectivo de identificar as classes homogéneas existentes no conjunto de dados.

A primeira parte do método de classificação consiste na aplicação do algoritmo de coloração sequencial ao grafo, definido sobre o conjunto a classificar, obtendo-se k_1 classes de cor. A segunda parte do método, consiste num ajuste da partição em k_1 classes de cor, resultando em $k \leq k_1$ classes homogéneas no contexto da análise classificatória.

O processo de ajuste consiste nos passos descritos a seguir. Para cada elemento u pertencente à classe de cor C_i com n_i elementos, seja $\tilde{n}_j$ o número de elementos da classe C_j pertencentes à bola com centro em u e raio α, $j = 1, \ldots, k_1$, $j \neq i$. Seja $\tilde{n}_m$ o máximo dos $\tilde{n}_j$. Se $\tilde{n}_m > n_i$ então o elemento u, assim como os restantes $n_i - 1$ elementos da classe C_i, são transferidos para a classe à qual os $\tilde{n}_m$ elementos pertencem.

No processo de ajuste algumas classes são "engolidas" por outras, fazendo com que alguns vértices adjacentes passem a pertencer à mesma classe. Vejamos de um modo esquemático como funciona o processo de ajuste:

1. Suponhamos que existem k_1 classes de cor designadas por $C_1, C_2, \ldots, C_{k_1}$.
2. $\forall i \in \{1, \ldots, k_1\}$ e $\forall v \in C_i$, seja $B(v, \alpha) \subset \Omega$, a bola centrada em v e de raio α.
3. Seja $\Pi_j(v) = C_j \cap B(v, \alpha)$ e $\tilde{n}_j(v) = Card\{\Pi_j(v)\}$, para $j = 1, \ldots, k_1$ $(j \neq i)$, ou seja, $\Pi_j(v)$ contém os elementos da classe C_j que estão a uma dissemelhança menor do que α de v.
4. Seja C_m a classe com maior número de elementos em $B(v, \alpha)$, ou seja,

$$\tilde{n}_m(v) = max_{\{j=1,\ldots,k_1\}}\{\tilde{n}_j(v)\}, j \neq i$$

5. Se $\tilde{n}_m(v) > n_i$, ou seja, se o elemento v está próximo de um maior número de elementos de uma classe diferente da classe à qual ele pertence, então o elemento v, assim como todos os elementos da classe de cor C_i, são transferidos para a classe m, isto é $C_m \leftarrow C_m \cup C_i$.

6. São retirados os arcos que uniam vértices de classes de cor diferentes e que agora pertencem à mesma classe.

Na figura 4.1 o elemento v, assim como os restantes elementos da mesma classe, são inseridos na classe 3, uma vez que é esta classe que tem maior número de elementos pertencentes à bola $B(v, \alpha)$, e superior ao número de elementos da classe à qual v pertence. Tem-se neste caso, para o elemento v, $\tilde{n}_1 = 4; \tilde{n}_2 = 5; \tilde{n}_3 = 6$ logo $\Pi_m(v) = \Pi_3(v)$ e os elementos que formavam uma classe independente, juntam-se aos elementos da classe 3.

Por esta abordagem, alguns vértices que pertenciam a classes de cor diferentes antes do processo de ajuste, podem tornar-se membros da mesma classe após este processo. No entanto, para que o método possa ter um suporte teórico no contexto da teoria dos grafos, pretende-se que se mantenha a propriedade de que só elementos não adjacentes pertencem à mesma classe. Para o efeito, retiram-se do grafo os arcos que unem vértices pertencentes à mesma classe (etapa (6)). Sejam C_i, $i = 1, \ldots, k$ as k classes resultantes do método e seja n_i o número de elementos da classe C_i, $i = 1, \ldots, k$. Neste caso verifica-se:

$$v_j, v_\ell \in C_i \Rightarrow \{v_j, v_\ell\} \notin E(G) \tag{4.14}$$

no entanto, e ao contrário do que acontece com as classes de cor,

$$v_j, v_\ell \in C_i \not\Rightarrow dist(v_j, v_\ell) < \alpha \tag{4.15}$$

permitindo assim a identificação de classes de forma alongada.

O processo de ajuste executado após a obtenção das classes de cor, não tem como objectivo reduzir ao mínimo o número de cores, no contexto da coloração em teoria de grafos. No caso do presente trabalho, o objectivo consiste em melhorar o desempenho do método na identificação das classes homogéneas, podendo uma situação óptima corresponder a um número de cores superior ao número cromático. Concluindo, podemos dizer que o algoritmo de coloração sequencial é executado puramente no contexto da teoria de grafos, sendo o ajuste realizado no contexto da análise classificatória.

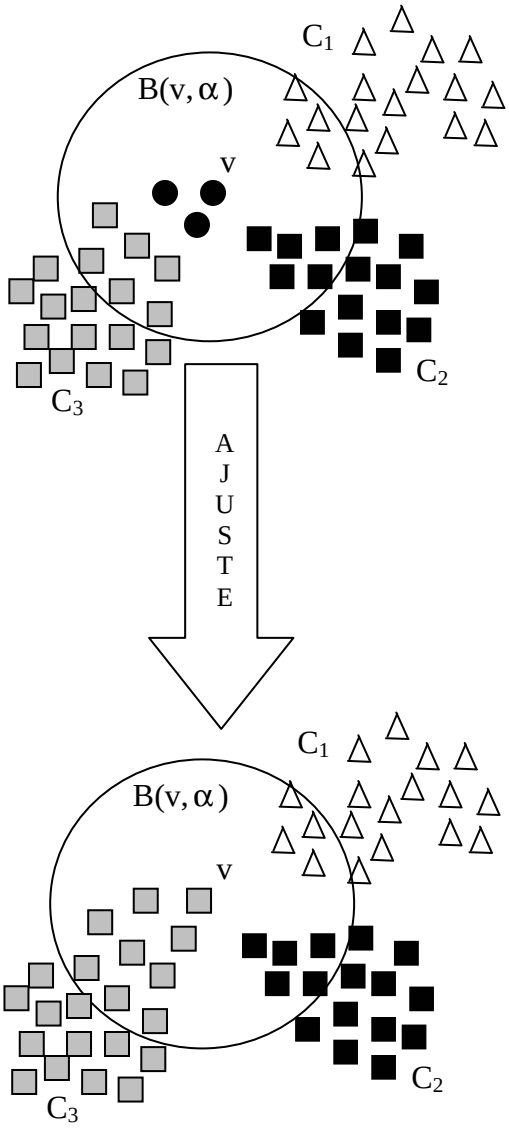

Figura 4.1: Exemplificação do processo de ajuste do algoritmo MPCG

.

Da aplicação de um algoritmo de coloração ajustado ao grafo representativo do conjunto de dados, resultam k classes com a propriedade de que vértices adjacentes não pertencem à mesma classe. Define-se, deste modo, um grafo *k-partite* no conjunto de dados dependente do parâmetro de controlo α. Este grafo é constituído por k classes de vértices não adjacentes, sendo *completo* se todos os vértices pertencentes a classes diferentes forem adjacentes.

4.3 Índice de Classificabilidade

Com o objectivo de validar a partição obtida pela aplicação do algoritmo MPCG, define-se em seguida um índice que avalia a classificabilidade de um conjunto de dados.

Associado a cada valor do parâmetro de controlo α, existe um grafo G definido sobre Ω. Da aplicação do método de partição ajustado ao grafo, resulta um grafo *k-partite*, no qual vértices pertencentes a classes diferentes podem ou não ser adjacentes. O índice descrito nesta secção, pretende identificar a partição que melhor representa a estrutura em classes do conjunto de dados, como sendo aquela que corresponde a um grafo *k-partite* completo. Assim, a identificação de um grafo *k-partite* completo no conjunto de dados, após a aplicação do algoritmo, implica a identificação de classes de vértices não adjacentes. A obtenção de um grafo *k-partite* completo depende do algoritmo de coloração usado, da ordem de entrada dos elementos e da estrutura dos dados. Em consequência, nem sempre se obtém, como resultado da aplicação de um algoritmo de coloração num grafo, um grafo *k-partite* completo.

O *índice de classificabilidade* orienta a procura da melhor partição, analisando o grafo *k-partite* que resulta de uma *k-coloração* ajustada dos vértices do grafo associado ao conjunto de dados. Como é sempre possível obter uma coloração de um grafo, existe sempre um grafo *k-partite* definido no conjunto de dados. O *Índice de Classificabilidade* (IC) mede o número de arcos em falta para que um grafo *k-partite* seja completo. O valor do índice é obtido somando para cada vértice, a diferença entre o seu grau máximo e o seu grau efectivo.

Sejam $C_1, \ldots, C_k$ as k classes, de $n_1, \ldots, n_k$ elementos respectivamente, obtidas pelo algoritmo MPCG e que caracterizam o grafo *k-partite* obtido.

Definição 4.11 *O valor do* Índice de Classificabilidade IC*, é dado por*

$$IC := \frac{1}{2}\sum_{i=1}^{k}\sum_{v\in C_i}[d_G^{max}(v) - d_G(v)] \tag{4.16}$$

sendo $d_G^{max}(v)$ *o grau máximo do vértice* v *e* $d_G(v)$ *o grau do vértice* v*.*

Variando o valor do parâmetro de controlo, o índice de classificabilidade identifica como a melhor partição, aquela que corresponde a um grafo com o menor número de arcos em falta para que seja *k-partite* completo.

Para cada valor fixo do parâmetro de controlo, o método determina a partição e o valor do índice IC associado. Escolhe-se a partição que melhor reflecte a estrutura dos dados, como aquela que corresponde ao melhor valor do índice.

Usando a propriedade que estabelece que num grafo *k-partite*, o grau máximo de um vértice pertencente a um conjunto com n_i vértices é $n-n_i$, e a equação (4.3), obtém-se:

$$\begin{aligned} IC &= \frac{1}{2}\sum_{i=1}^{k}\sum_{v\in C_i}[d_G^{max}(v) - d_G(v)] = \frac{1}{2}\left(\sum_{i=1}^{k}\sum_{v\in C_i}(n-n_i) - \sum_{i=1}^{k}\sum_{v\in C_i} d_G(v)\right) \\ &= \frac{1}{2}\left(\sum_{i=1}^{k} n_i(n-n_i) - 2m\right) = \frac{1}{2}\left(\sum_{i=1}^{k} n_i n - \sum_{i=1}^{k} {n_i}^2 - 2m\right) \\ &= \frac{1}{2}\left(n^2 - \sum_{i=1}^{k} n_i^2 - 2m\right) = \frac{1}{2}\left(n^2 - \sum_{i=1}^{k} n_i^2\right) - m \end{aligned} \tag{4.17}$$

onde m é o número de arcos no grafo.

O índice é normalizdo pelo número máximo de arcos no grafo *k-partite* completo, dado por

$$E_{max} = \frac{1}{2}\sum_{i=1}^{k} n_i(n-n_i) = \frac{1}{2}\left(n^2 - \sum_{i=1}^{k} n_i^2\right) \tag{4.18}$$

que corresponde à soma dos graus máximos de todos os vértices do grafo definido nos dados. Assim, o *índice de classificabilidade normalizado* é dado por,

$$\tilde{IC} := \frac{E_{max} - m}{E_{max}} = 1 - \frac{m}{E_{max}} \tag{4.19}$$

No caso extremo de todos os vértices distarem entre si menos do que α, não existem vértices adjacentes, e consequentemente, o grafo não tem arcos. Neste caso, todos os vértices podem ser coloridos com uma só cor e o conjunto de dados é constituído por uma só classe. Assim, verifica-se $E_{max} = 0$, porque o grau máximo de cada vértice é zero. Neste caso, o valor de $\tilde{IC}$ é indeterminado, optando-se pela atribuição do valor -1. Por outro lado, se todos os vértices distarem entre si mais do que α, todos os vértices são adjacentes, e são necessárias n cores para se obter uma coloração do grafo. Assim, existem n classes de cor com um só elemento cada, e neste caso, $\tilde{IC} = 0$.

Para grafos não completos, se $\tilde{IC} = 0$ tem-se o caso ideal em que o conjunto de dados é composto por k classes compactas e isoladas. No entanto, uma partição localmente óptima pode corresponder a um valor do índice de classificabilidade diferente de zero, mas que corresponde a um decréscimo do valor do índice comparando com valores anteriores.

4.4 Parâmetro de Controlo

A eficácia do método na identificação da estrutura em classes homogéneas do conjunto de dados, depende do valor do parâmetro de controlo α. O valor deste parâmetro determina o número de arcos no grafo e consequentemente, define as adjacências entre os vértices, sendo estas determinantes na identificação do grafo *k-partite* no conjunto de dados. Em consequência, depende de α o número de classes do conjunto de dados, assim como o número de elementos em cada uma das classes, quantidades das quais depende o índice de classificabilidade. Assim, para diferentes valores do parâmetro de controlo, obtemos diferentes estruturas em classes identificadas pelo método ajustado. Na implementação proposta, este parâmetro pode por opção, não ser fornecido pelo utilizador, sendo-lhe automaticamente atribuídos alguns valores.

Seja d_{min} a dissemelhança mínima e d_{max} a dissemelhança máxima, entre pares de elementos do conjunto de dados. De acordo com a definição do grafo no conjunto de dados (Secção 4.2.1), dois vértices são adjacentes no grafo G, se a sua dissemelhança for maior ou igual a α. Assim, valores do parâmetro próximos de d_{min} correspondem a

um maior número de arcos no grafo e, por consequência, a um maior número de classes até um máximo de $k = n$ (onde n é o número de elementos do conjunto de dados). Por outro lado, valores do parâmetro próximos de d_{max} correspondem a um menor número de arcos e a um menor número de classes, até um mínimo de $k = 1$ para $\alpha = d_{max}$. O valor do parâmetro α pode variar entre o limite mínimo e o limite máximo sendo, para cada valor, calculado o índice de classificabilidade que pretende identificar as classes homogéneas que melhor reflectem a estrutura do conjunto de dados.

Na implementação efectuada, os valores que pode assumir o parâmetro de controlo α são decididos pelo utilizador. Se existir algum conhecimento sobre os dados, o valor do parâmetro pode ser atribuído directamente pelo utilizador. No entanto, uma execução automática do método permite que os valores de α sejam atribuídos de acordo com opções feitas pelo utilizador.

Assim, o valor do parâmetro de controle α pode tomar os seguintes valores:

- α é dado pelo utilizador.
- α toma o valor da média das dissemelhanças entre todos os pares de elementos de conjunto de dados, isto é :

$$\alpha = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d(x_i, x_j) \tag{4.20}$$

- α é tal que $\ell_{min} \leq \alpha \leq \ell_{max}$ ($\ell_{min} \geq d_{min}$ e $\ell_{max} \leq d_{max}$), sendo o intervalo $[\ell_{min}, \ell_{max}]$ dividido em H partes, e o método executado para cada $\alpha = \ell_{min} + ih$, para $i = 1, \ldots, H$ e $h = \dfrac{\ell_{max} - \ell_{min}}{H}$. Os valores de ℓ_{min}, ℓ_{max} e H são seleccionados pelo utilizador, podendo ser, por opção deste, $\ell_{min} = d_{min}$, $\ell_{max} = d_{max}$. De entre as H partições, é seleccionada aquela que corresponde ao melhor valor do índice $\tilde{IC}$.

4.5 Aplicação do Método

Nesta secção, o método ajustado e o índice de classificabilidade são aplicados a quatro conjuntos de dados caracterizados por dois atributos numéricos. Foi usada a distância

Tabela 4.1: Parâmetros do gerador de dados para as elipses do conjunto de dados representado na figura 4.3

p	a	b	a_1	b_1	a_2	b_2	a_3	b_3	p_1	p_2	p_3
25	100	50	30	10	30	20	40	30	100	0	0
25	100	90	30	10	30	20	40	30	100	0	0
25	100	130	30	10	30	20	40	30	100	0	0

euclideana aplicada aos dados não estandardizados.

Os conjuntos de dados foram gerados pelo gerador de dados artificiais apresentado na Secção 3.10.1.

4.5.1 Conjunto de Dados com Classes de forma Alongada

O primeiro conjunto de dados considerado tem 2000 elementos distribuídos por quatro classes e está representado na figura 4.3. As tabelas 4.1 e 4.2 indicam os parâmetros do gerador de dados para as três elipses e para o círculo do conjunto de dados.

A figura 4.2 representa os valores do índice $\tilde{IC}$ e o número de classes, resultantes da aplicação do método aos dados, para diferentes valores do parâmetro de controlo α, onde $d_{min} \leq \alpha \leq d_{max}$ e $H = 50$. Pela análise do grafo conclui-se que o valor de $\tilde{IC}$ identificativo do número de classes é $\tilde{IC} = 0.005275$ correspondente a uma partição em quatro classes, obtida para $\alpha = 21.77236$ e representada na figura 4.3. Os valores para os quais $\tilde{IC} = -1$ não foram incluídos no gráfico.

Foram gerados 100 conjuntos de dados e as suas partições de referência com os parâmetros indicados nas tabelas 4.1 e 4.2. O método foi aplicado a cada conjunto de dados e a melhor partição obtida foi comparada com a partição de referência usando o índice de *Rand* corrigido de *Hubert & Arabie* (IRC) (Secção 2.5.5). A média e o desvio padrão dos valores do índice IRC obtidos para os 100 conjuntos foram $\bar{x} = 0.989$ e $s = 0.038$. Devido à geometria deste conjunto de dados conclui-se que para este exemplo, o método identificou com sucesso classes de forma alongada.

Tabela 4.2: Parâmetros do gerador de dados para o círculo do conjunto de dados representado na figura 4.3

p	a	b	r_1	r_2	r_3	p_1	p_2	p_3
25	50	100	10	30	40	100	0	0

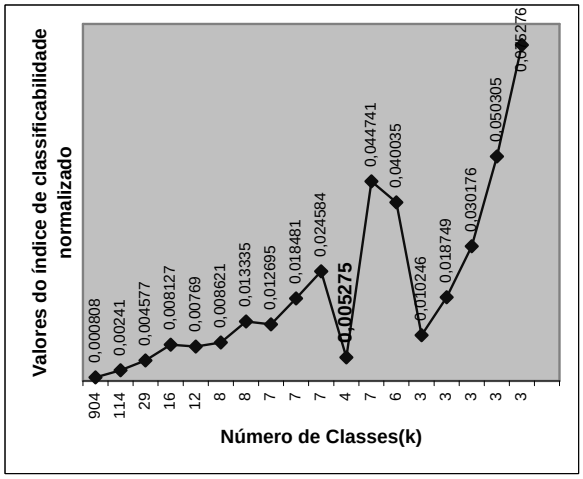

Figura 4.2: Gráfico $\tilde{IC}/k$ obtido para o conjunto de dados com classes alongadas.

4.5.2 Conjunto de Dados com Pontos Isolados

O segundo conjunto de dados considerado tem 1000 elementos, sendo 900 distribuídos por cinco classes compactas e isoladas, enquanto os restantes 100 elementos são uniformemente distribuídos pela janela na qual os elementos estão localizados. A tabela 4.3 indica os parâmetros do gerador de dados para as cinco classes deste conjunto de dados representado na figura 4.5.

A figura 4.4 representa os valores do índice $\tilde{IC}$ e o número de classes, resultantes

Tabela 4.3: Parâmetros do gerador de dados para as cinco classes representadas na figura 4.5

p	a	b	r_1	r_2	r_3	p_1	p_2	p_3
20	50	50	20	20	20	100	0	0
20	200	50	20	20	20	100	0	0
20	50	200	20	20	20	100	0	0
20	125	125	20	20	20	100	0	0
20	200	200	20	20	20	100	0	0

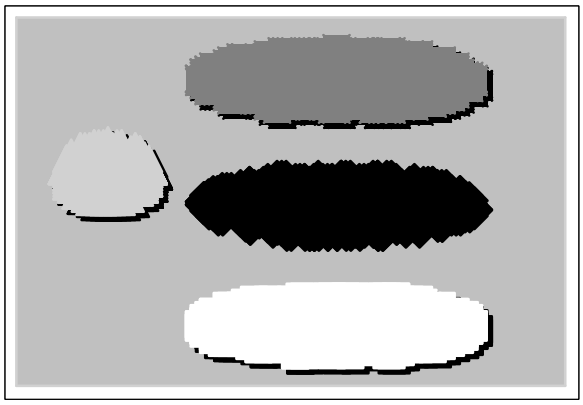

Figura 4.3: Partição em quatro classes obtida para o conjunto de dados com classes alongadas.

da aplicação do método ao conjunto de dados para diferentes valores do parâmetro de controlo α, onde $d_{min} \leq \alpha \leq d_{max}$ e $H = 50$. Pela análise do gráfico conclui-se que o valor óptimo é $\tilde{IC} = 0.000044$, correspondendo a uma partição em 41 classes, obtidas para $\alpha = 23.529103$. Como mostra a figura 4.5, que representa a partição em 41 classes obtidas pelo método, os pontos isolados não mascararam a distribuição dos elementos pelas cinco classes. Nesta partição, 929 elementos são distribuídos pelas cinco classes, sendo os pontos isolados distribuídos por uma classe com 5 elementos, quatro classes com 4 elementos, quatro classes com 3 elementos, dez classes com 2 elementos e dezoito classes com um elemento.

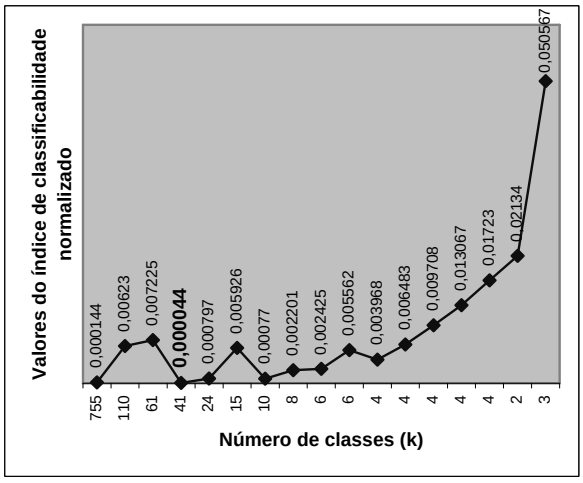

Figura 4.4: Gráfico $\tilde{IC}/k$ obtido para o conjunto de dados com pontos isolados.

Foram gerados 100 conjuntos de dados e as suas partições de referência para os

parâmetros indicados na tabela 4.3. A partição de referência tem seis classes, tendo a sexta classe todos os pontos isolados com distribuição uniforme. O método foi aplicado a cada conjunto de dados e a melhor partição foi comparada com a partição de referência usando o índice IRC. A média e o desvio padrão dos valores do índice IRC obtidos para os 100 conjuntos foram $\bar{x} = 0.922$ e $s = 0.077$. Devido à geometria do conjunto de dados, pode-se concluir que para este exemplo, o método é robusto relativamente a pontos isolados.

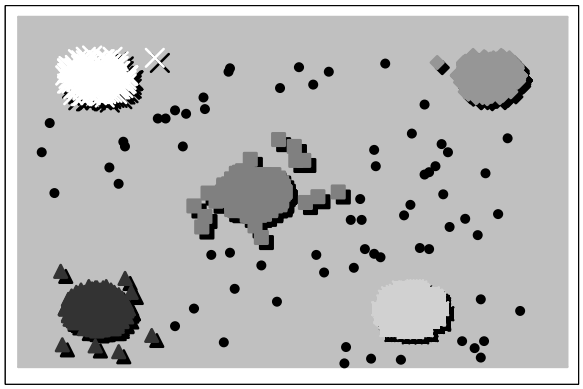

Figura 4.5: Partição em 41 classes obtida para o conjunto de dados com pontos isolados.

4.5.3 Conjunto de Dados com Classes de Forma Não Convexa

O terceiro exemplo consiste num conjunto de dados com 500 elementos, distribuídos por duas classes compactas e isoladas com formas não convexas [1] tal como representado na figura 4.7. A tabela 4.4 indica os parâmetros do gerador de dados para as duas classes deste conjunto. Para este caso, o gerador de dados foi instruído para gerar só meios arcos.

A figura 4.6 representa os valores do índice $\tilde{IC}$ e o número de classes, resultantes da aplicação do método ao conjunto de dados para diferentes valores do parâmetro de controlo α, onde $d_{min} \leq \alpha \leq d_{max}$ e $H = 50$. Obtém-se a partição com 2 classes que melhor reflectem a estrutura dos dados, para $\alpha = 58.937767$ e $\tilde{IC} = 0.000048$, como pode ser visto na figura 4.7.

[1] Uma versão deste exemplo com 400 elementos foi usado em [72] para testar o algoritmo *voting K-means*.

Tabela 4.4: Parâmetros do gerador de dados para as classes do conjunto representado na figura 4.7

p	a	b	r_1	r_2	r_3	p_1	p_2	p_3
50	100	110	80	100	100	0	100	0
50	190	90	80	100	100	0	100	0

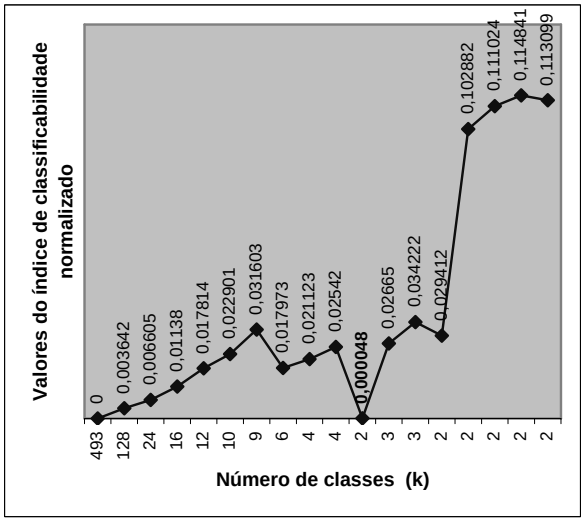

Figura 4.6: Gráfico $\tilde{IC}/k$ obtido para o conjunto de dados com classes de forma não convexa.

Foram gerados 100 conjuntos de dados e as suas partições de referência para os parâmetros indicados na tabela 4.4. O método foi aplicado a cada conjunto de dados e a melhor partição obtida foi comparada com a partição de referência usando o *Índice de Rand Corrigido de Hubert e Arabie* (IRC). A média e o desvio padrão dos valores do índice IRC obtidos foram $\bar{x} = 0.942$ e $s = 0.089$. Devido à geometria do conjunto de dados podemos concluir que para este exemplo, o método identificou com sucesso classes de forma não convexa.

4.5.4 Conjunto de Dados sem Estrutura em Classes

A figura 4.8 mostra o gráfico $\tilde{IC}/k$, resultado da aplicação do método para diferentes valores do parâmetro de controlo α, onde $d_{min} \leq \alpha \leq d_{max}$ e $H = 50$, a um conjunto de dados com 1000 elementos com uma distribuição Uniforme na janela unitária. Como se pode concluir da análise deste gráfico, os valores de $\tilde{IC}$ para diferentes valores de α, não nos permitem concluir àcerca do valor indicativo do número de classes. Este

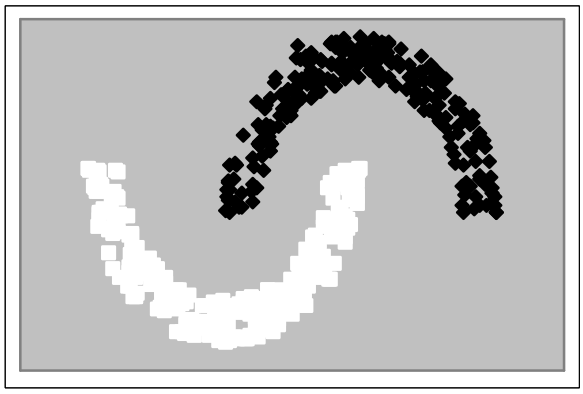

Figura 4.7: Partição em duas classes obtida para o conjunto de dados com classes de forma não convexa.

facto indica a ausência de uma estrutura em classes no conjunto de dados.

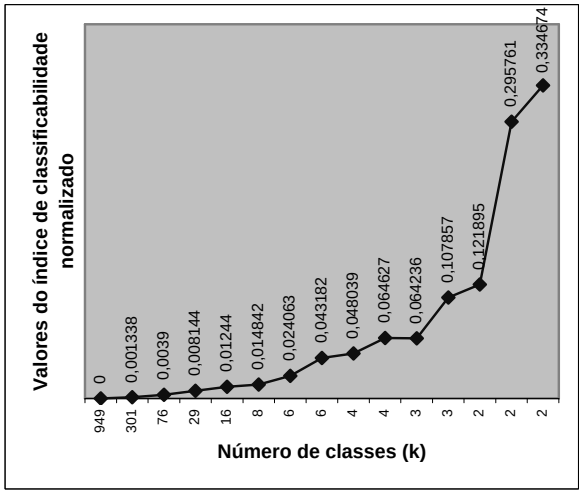

Figura 4.8: Gráfico $I\tilde{C}/k$ obtido para o conjunto de dados sem estrutura em classes.

Foram gerados 100 conjuntos de dados com uma distribuição Uniforme, cada um dos gráficos correspondentes a $I\tilde{C}$ está representado na figura 4.9. Como se pode ver na figura, todas as curvas são semelhantes, impossibilitando a escolha do valor de $I\tilde{C}$ indicativo do número de classes. O valor médio da diferença máxima entre os valores de $I\tilde{C}$ entre quaisquer duas curvas é 0.060758. Este exemplo mostra que o método proposto tratou com sucesso o problema de detectar um conjunto de dados sem nenhuma estrutura em classes.

A tabela 4.5 representa o resumo dos valores do índice IRC obtidos para os três

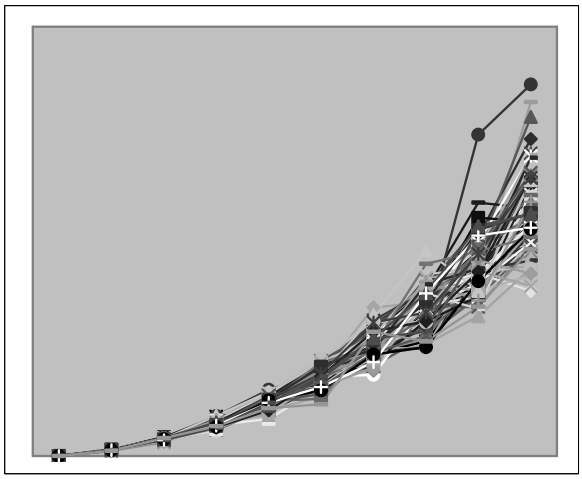

Figura 4.9: Gráfico $\tilde{IC}$ obtido para 100 conjuntos de dados sem estrutura em classes.

Tabela 4.5: Média e desvio padrão dos valores do índice IRC de comparação com a partição de referência, para 100 execuções dos três conjuntos de dados testados.

Conjuntos de Dados	Média	Desvio Padrão
Classes de forma alongada	0.989	0.038
Pontos isolados	0.922	0.077
Classes de forma não convexa	0.942	0.089

primeiros exemplos.

4.6 Estudo da Estabilidade do Método

Nesta secção pretende-se avaliar a estabilidade do método de classificação, reanalisando uma versão modificada do conjunto de dados e avaliando a extensão das diferenças, em relação à classificação original. O objectivo consiste em isolar e analisar aspectos específicos do desempenho do método de classificação, tais como a obtenção da estrutura inerente nos dados, a sua sensibilidade quanto à reorganização dos dados e a estabilidade dos resultados perante novos dados, dados retirados ou dados perturbados.

O conjunto de dados artificial estudado é constituído por 1000 elementos, caracterizados por dois atributos numéricos. As tabelas 4.6 e 4.7 indicam os parâmetros do gerador de dados para as quatro classes esféricas e para a classe elíptica do conjunto de dados, representado na figura 4.11. O gráfico da figura 4.10 mostra os resultados

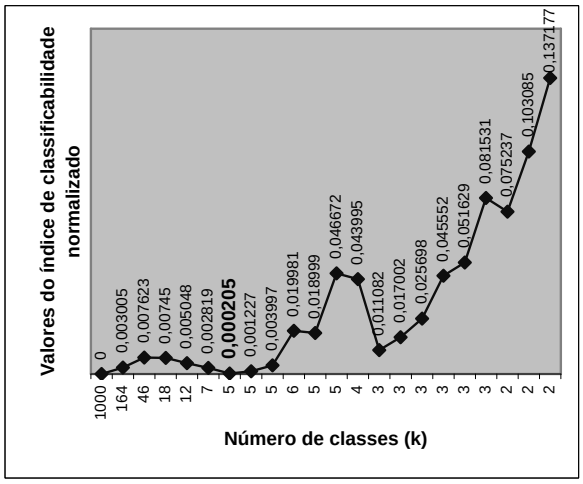

Figura 4.10: Gráfico $\tilde{IC}/k$ obtido para um conjunto de dados com 1000 elementos e cinco classes.

Tabela 4.6: Parâmetros do gerador de dados para as classes esféricas do conjunto de dados representado na figura 4.11.

p	a	b	r_1	r_2	r_3	p_1	p_2	p_3
20	50	170	20	20	20	100	0	0
20	80	230	20	20	20	100	0	0
20	150	50	20	30	50	85	10	5
20	200	150	30	40	50	80	10	10

obtidos pelo método, para cada um dos valores do parâmetro de controlo, onde $d_{min} \leq \alpha \leq d_{max}$ e $H = 50$. Pela leitura do gráfico conclui-se que a melhor partição é obtida para $\tilde{IC} = 0.000205$ e $k = 5$, valores obtidos quando $\alpha = 32.313587$. A figura 4.11 representa a partição identificada pelo método. Consideremos esta partição como a partição de referência.

Seguem-se algumas experiências sugeridas em vários trabalhos encontrados na literatura ([91], [164], [159] e [189]) e que permitem avaliar alguns aspectos importantes do desempenho do método.

Tabela 4.7: Parâmetros do gerador de dados para a classe elíptica do conjunto de dados representado na figura 4.11.

p	a	b	a_1	b_1	a_2	b_2	a_3	b_3	p_1	p_2	p_3
20	50	100	50	15	50	15	55	20	90	0	10

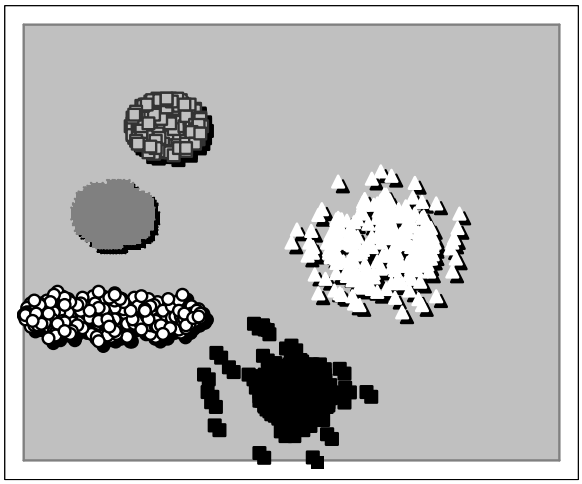

Figura 4.11: Distribuição em cinco classes obtida para o conjunto de dados cujo gráfico $I\tilde{C}/k$ está representado na figura 4.10.

1. *Permutação dos dados de entrada.* Um dos factores que condicionam o desempenho do método de classificação *Método de Partição baseado na Coloração de Grafos* (MPCG), é a ordem pela qual os elementos são classificados. Assim, nesta experiência pretende-se avaliar a alteração sofrida no desempenho do método na identificação da estrutura dos dados, fazendo permutações no conjunto de dados. O método foi aplicado 100 vezes ao conjunto de dados fazendo em cada execução uma permutação aleatória dos elementos, assegurando assim uma ordem diferente em cada execução. A melhor partição obtida para cada conjunto foi comparada com a partição de referência usando o *Índice de Rand Corrigido de Hubert e Arabie* (IRC) (Secção 2.5.5). A média e o desvio padrão dos valores do índice IRC obtidos para as 100 execuções foram $\bar{x} = 0.977$ e $s = 0.058$.

   Para os conjuntos de dados testados, podemos concluir que o método é robusto quanto à ordem de entrada dos elementos a classificar.

2. *Retirar* 10% *dos elementos.* Foram retirados de modo aleatório 100 elementos do conjunto de dados; aplicou-se o método ao conjunto de dados assim reduzido, sendo a melhor partição obtida, comparada com a partição de referência usando o índice IRC. A experiência foi repetida 100 vezes, sendo a média e o desvio padrão dos valores do índice IRC obtidos, $\bar{x} = 0.984$ e $s = 0.050$.

   Para os conjuntos de dados testados, o método mostrou-se robusto relativamente à ausência de alguns elementos.

3. *Introduzir* 10% *de pontos isolados.* Para este caso considerou-se a experiência já realizada na Secção 4.5.2, na qual foram gerados 100 conjuntos de dados com 900 elementos distribuídos por cinco classes mais 100 elementos distribuídos uniformemente. A média e o desvio padrão dos índices IRC obtidos foram $\bar{x} = 0.922$ e $s = 0.077$.

   Para os conjuntos de dados considerados na secção referida, o método mostrou ser robusto face à presença de pontos isolados. O método, além de não ser negativamente influenciado na identificação das cinco classes principais, pela presença de pontos isolados, ainda consegue identificar a maior parte dos pontos isolados.

4. *Perturbar as distâncias euclideanas.* As distâncias euclideanas entre elementos do conjunto são perturbadas, sendo recalculadas como $\sum_{j=1}^{t}(x_{ij} - x_{\ell j} - e_{i\ell j})^2$ sendo x_{ij} o valor da coordenada j do elemento x_i, $x_{\ell j}$ o valor da coordenada j do elemento x_ℓ e $e_{i\ell j}$ é o valor da perturbação feita na distância, para a variável j, dos elementos x_i e x_ℓ. Os valores de $e_{i\ell j}$ são obtidos a partir de um gerador de números aleatórios com distribuição univariada Normal com média zero e desvio padrão igual à média dos desvios padrão na dimensão j para as duas classes que contêm os elementos x_i e x_ℓ [159]. A experiência foi repetida 100 vezes, sendo a média e o desvio padrão dos valores do índice IRC obtidos $\bar{x} = 0.888$ e $s = 0.138$.

   Com a perturbação das distâncias entre os elementos do conjunto de dados, verifica-se que em alguns casos, algumas classes não ficaram totalmente separadas, tendo o método, nestes casos, maior dificuldade em identificá-las como classes separadas. Esta situação verificou-se em 33% dos conjuntos perturbados.

5. *Aumentar a dimensão com variáveis de valores uniformes em* $[0, 1]$. Foi introduzida uma variável cujos valores numéricos foram obtidos por um gerador de números aleatórios no intervalo $[0, 1]$. As outras variáveis foram estandardizadas. Seguiu-se a aplicação do método ao conjunto de dados de dimensão três, sendo a partição obtida, comparada com a de referência usando o IRC. A experiência foi repetida 100 vezes, sendo a média e o desvio padrão dos valores do índice IRC obtidos, $\bar{x} = 0.978$ e $s = 0.098$.

   O resultado mostra que a introdução de uma variável "ruído" perturbou pouco a identificação das classes.

A tabela 4.8 representa os valores do índice IRC obtidos para 100 execuções dos cinco

Tabela 4.8: Média e desvio padrão dos valores do índice IRC, para 100 execuções de todos os conjuntos de dados testados.

Perturbação dos Dados	Média	Desvio Padrão
Permutar os elementos	0.977	0.058
Retirar 10% dos elementos	0.984	0.050
Introduzir 10% pontos isolados	0.922	0.077
Perturbar as distâncias	0.888	0.138
Aumentar a dimensão	0.978	0.098

casos de perturbação dos dados testados.

4.7 Diferentes Abordagens para a Coloração de Vértices

O algoritmo de coloração sequencial clássico depende da ordem de entrada dos dados e faz a atribuição de cores pela ordem pela qual os vértices se apresentam. Designemos o método de classificação que aplica este algoritmo por *método de partição baseado na coloração de grafos ajustado usando ordem inicial*, MPCG_OI. No entanto, existem outros algoritmos sequenciais que fazem uma selecção diferente dos vértices a colorir, de modo a minimizar o número de cores usadas na coloração. Por exemplo, o algoritmo *maior grau primeiro* selecciona os vértices a colorir por ordem decrescente dos seus graus [220]; este algoritmo apresenta um fraco desempenho em grafos *bi-partite* [45], ganhando um grande incremento de eficácia e eficiência na sua versão em paralelo [59]. O algoritmo *ordenação de graus de saturação* selecciona para vértice a colorir aquele que apresenta o maior número de cores adjacentes diferentes [34]; este algoritmo tem o grande inconveniente de o tempo de execução e espaço utilizado serem excessivos [45], sendo o tempo de execução proporcional a $\sum_{v \in V} d_G^2(v)$ ([34] e [85]). Uma outra alternativa, para a ordem pela qual os vértices são coloridos, consiste numa escolha dos vértices seguindo uma *ordenação de incidência de graus* ([45] e [123]). Seguindo este critério, suponhamos que os vértices $v_1, \ldots, v_{i-1}$ já foram coloridos. O vértice v_i é escolhido de entre todos os não coloridos, como o que apresenta o maior grau no sub-grafo de G induzido pelo conjunto $\{v_1, \ldots, v_{i-1}\} \cup \{v_i\}$. Designemos o método de classificação que aplica este algoritmo por *método de partição baseado na coloração de grafos ajustado usando ordenação de incidência de graus*, MPCG_OIG. Para diminuir o tempo de execução desta abordagem, note-se que se os vértices u e v são não adjacentes em G, então estes vértices podem ser coloridos em paralelo [123].

Uma outra abordagem, no contexto da coloração de vértices, baseia-se na procura de conjuntos independentes maximais no conjunto de dados. A analogia entre estes conjuntos e coloração óptima surge uma vez que se pretende encontrar o menor número possível de classes de cor, e assim, pretende-se que as classes tenham o maior número de vértices não adjacentes possível. Ou seja, pretende-se que C_0 seja um conjunto independente maximal em V, que C_1 seja um conjunto independente maximal em $V \setminus C_0$, etc. Este tipo de algoritmos aceita como dado de entrada o grafo $G(V, E)$, tendo como saída um conjunto independente maximal que corresponde a uma classe homogénea. Seja $G'(V', E')$ um sub-grafo de G. Para todo $W \subseteq V'$, seja $Viz(W)$ a vizinhança de W. O algoritmo segue, de modo muito genérico os seguintes passos [150]:

1. $I \leftarrow \emptyset$
2. $G' = (V', E') \leftarrow G = (V, E)$
3. enquanto $G' \neq \emptyset$ faça
   (a) seleccionar um conjunto $I' \subseteq V'$ que seja independente em G'
   (b) $I \leftarrow I \cup I'$
   (c) $Y \leftarrow I' \cup Viz\{I'\}$
   (d) o grafo $G' = (V', E')$ é o subgrafo induzido em $V' \setminus Y$

Quando o algoritmo termina, o conjunto I é um conjunto independente maximal. O passo (3a) é o que apresenta maior complexidade e maior consumo de tempo, sendo ainda reconhecido como sendo NP-completo. Em [150] é apresentada uma execução deste passo em paralelo, reduzindo assim significativamente o tempo de execução do mesmo.

No entanto, pode-se optar por uma solução não óptima de procura de conjuntos independentes seguindo o algoritmo:

1. $j = 0$
2. enquanto $V \neq \emptyset$
   (a) $V' \leftarrow V$

(b) enquanto $V' \neq \emptyset$

i. escolher vértice a colorir v

ii. $C_j \leftarrow C_j \cup \{v\}$

iii. $V' \leftarrow V' \setminus \{Viz\{v\} \cup \{v\}\}$

iv. $V \leftarrow V \setminus \{v\}$

(c) $j \leftarrow j + 1$

A escolha do vértice a colorir no passo 2(b)i pode ser executada de diferentes formas:

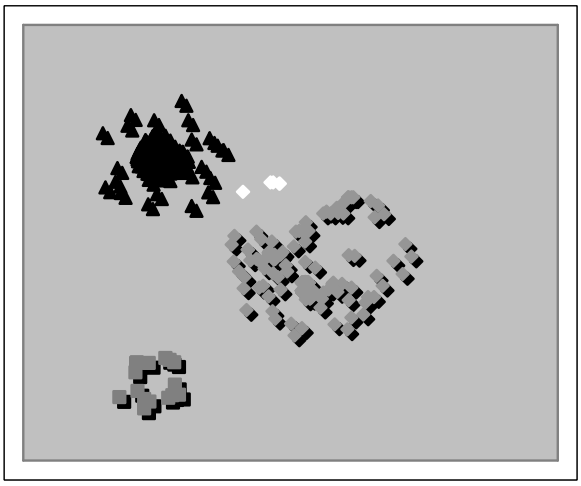

Figura 4.12: Distribuição em quatro classes do conjunto de dados com 400 elementos, identificada pelo método MPCG_CIM_MGP.

1. O vértice v a colorir resulta de uma escolha aleatória de entre os vértices ainda não coloridos [150]. Designemos o método de classificação que aplica este algoritmo por *método de partição baseado na coloração de grafos ajustado usando o conjunto independente maximal com selecção aleatória*, MPCG_CIM_SA.

2. O vértice v a colorir é escolhido sequencialmente pela ordem pela qual aparece no conjunto de vértices a colorir [151]. Designemos o método de classificação que aplica este algoritmo por *método de partição baseado na coloração de grafos ajustado usando o conjunto independente maximal com selecção do mínimo*, MPCG_CIM_SM.

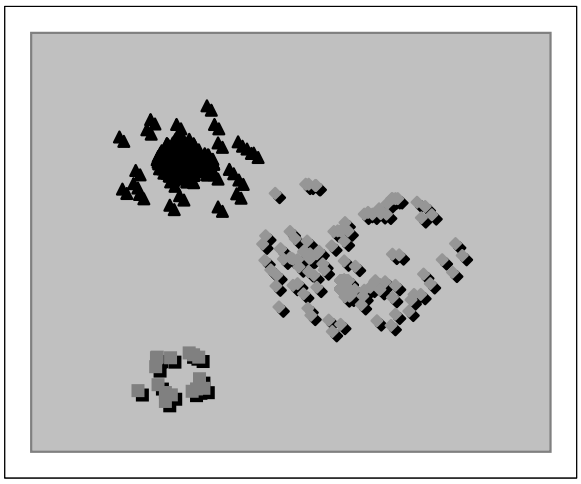

Figura 4.13: Distribuição em três classes do conjunto de dados com 400 elementos, identificada pelos métodos MPCG_OI, MPCG_OIG, MPCG_CIM_SA e MPCG_CIM_SM

3. O vértice v a colorir é seleccionado da seguinte maneira:

$$\begin{cases} \text{escolher o primeiro vértice ainda não colorido} \\ i \leftarrow max\left\{max_{\{x \in Viz\{v\}\}}\{d_G(x)\}; d_G(v)\right\} \\ v \leftarrow x : d_G(x) = i \end{cases}$$

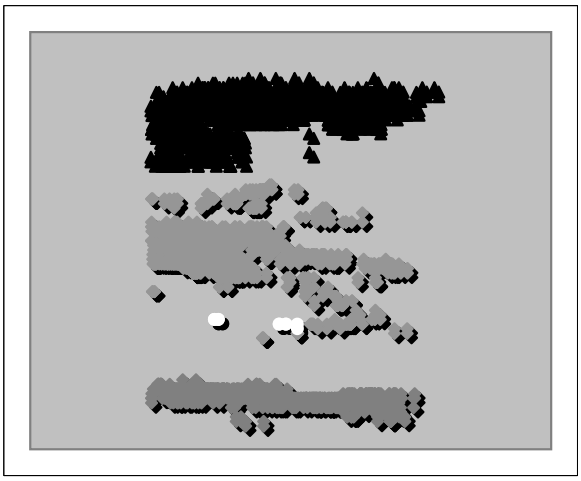

Figura 4.14: Distribuição em quatro classes do conjunto de dados com 3892 elementos, identificada pelos métodos MPCG_OI e MPCG_CIM_SM.

Figura 4.15: Distribuição em três classes do conjunto de dados com 3892 elementos, identificada pelos métodos MPCG_CIM_SA, MPCG_OIG e MPCG_OI.

Tabela 4.9: Resultados para diferentes inicializações do algoritmo de coloração.

	Conjunto 1		Conjunto 2		Conjunto 3	
	k	tempo (seg.)	k	tempo (seg.)	k	tempo (seg.)
MPCG_OI	3	1.27	4	43.91	4	660.09
MPCG_OIG	3	26.45	4	3300.7	3	24863.73
MPCG_CIM_SA	3	0.86	4	69.34	3	996.86
MPCG_CIM_SM	3	1.22	4	40.06	4	650.01
MPCG_CIM_MGP	4	12.11	4	533.98	4	5771.45

ou seja, escolhe-se sequencialmente um vértice candidato a ser colorido e compara-se o seu grau com os graus dos vértices que lhe são adjacentes, escolhendo-se para colorir o vértice de maior grau [59]. Designemos o método de classificação que aplica este algoritmo por *método de partição baseado na coloração de grafos ajustado usando o conjunto independente maximal com selecção do maior grau primeiro*, MPCG_CIM_MGP.

Estes métodos foram aplicados a três conjuntos de dados com 2 atributos numéricos. Foram analisados o número de classes identificadas assim como os tempos gastos, sendo a execução feita num PC *Pentium 4* de 3GHz a 512 Mbytes de RAM. O primeiro conjunto de dados contém 400 elementos e está representado na figura 4.12; o segundo conjunto já foi usado na Secção 4.5.1 e está representado na figura 4.3; o terceiro é um conjunto de dados reais com 3892 elementos, está representado na figura 4.15 e consiste em três partes da rede rodoviária Grega [103].

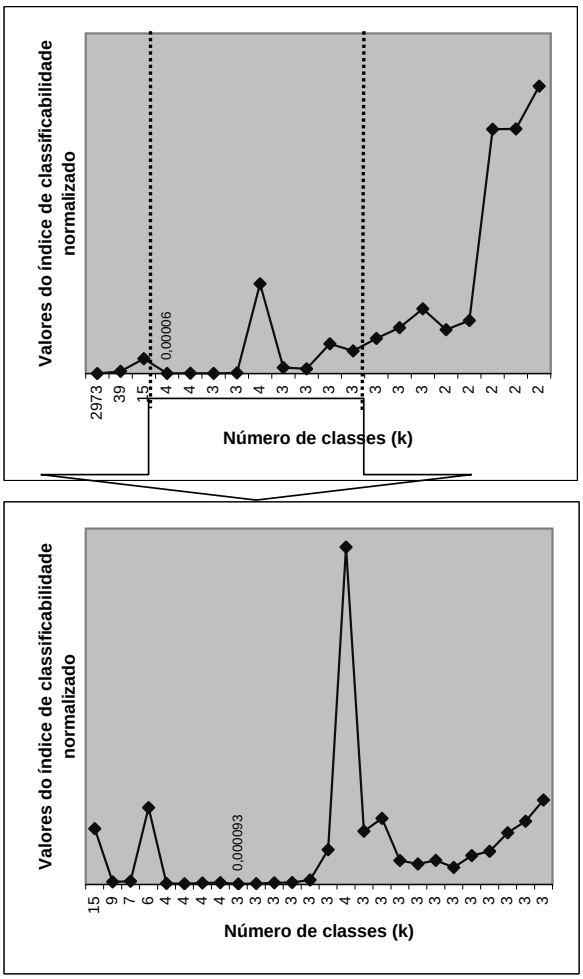

Figura 4.16: Gráfico $\tilde{IC}/k$ obtido para o conjunto de dados com 3892 elementos.

Os métodos MPCG_OI, MPCG_OIG, MPCG_CIM_SA e MPCG_CIM_SM identificaram a partição representada na figura 4.13, enquanto o método MPCG_CIM_MGP identificou a partição representada na figura 4.12. Pode-se concluir que os métodos testados identificaram com sucesso as classes pelas quais os elementos estão distribuídos, sendo o último o mais eficaz; no entanto, tal como pode ser consultado na tabela 4.9, este é o segundo método mais lento, só superado pelo MPCG_OIG, não havendo diferenças significativas de tempos entre os outros três métodos.

Quanto ao segundo exemplo, todos os métodos identificaram correctamente a partição em 4 classes representada na figura 4.3, no entanto, como se pode ver na tabela 4.9, as diferenças de tempo de execução entre métodos já são mais significativas, sendo os

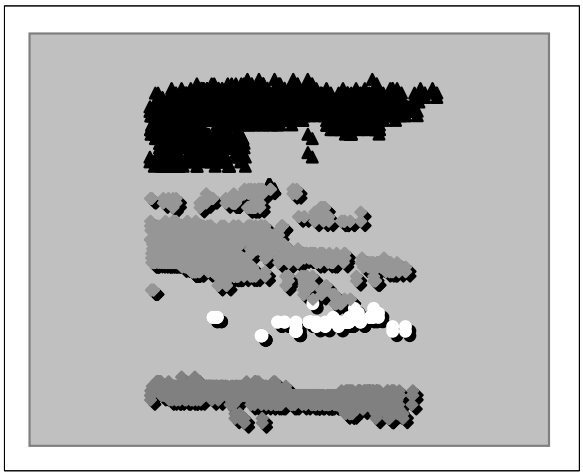

Figura 4.17: Distribuição em quatro classes do conjunto de dados com 3892 elementos, identificada pelo método MPCG_CIM_MGP.

métodos mais eficientes os MPCG_OI e MPCG_CIM_SM e o mais lento o MPCG_OIG, sendo aproximadamente 82 vezes mais lento que os dois mais eficientes.

No caso do terceiro exemplo com 3892 elementos, a figura 4.14 representa a partição identificada pelos métodos MPCG_OI e MPCG_CIM_SM; a partição obtida pelos métodos MPCG_CIM_SA e MPCG_OIG está representada na figura 4.15, enquanto a figura 4.17 representa a partição obtida pelo método MPCG_CIM_MGP. No caso do método MPCG_OI, obteve-se o gráfico representado na parte superior da figura 4.16 (para $\ell_{min} = d_{min}$, $\ell_{max} = d_{max}$ e $H = 50$) que identificou a melhor partição para $\alpha = 45152.58$, $\tilde{IC} = 0.00006$ e $k = 4$. No entanto, existem outros valores de k com valores de $\tilde{IC}$ muito próximos do valor óptimo. Assim, justifica-se uma nova execução do método restringindo o intervalo no qual toma valores o parâmetro de controlo α para valores próximos dos valores mínimos locais. A segunda execução fez-se para $\ell_{min} = 30102.06$, $\ell_{max} = 165556.8$ e $H = 25$. O gráfico respectivo encontra-se na parte inferior do figura 4.16, onde é identificada como melhor partição aquela que corresponde a $\alpha = 73447.58$, $\tilde{IC} = 0.000093$ e $k = 3$ e que corresponde à partição identificada na figura 4.15.

Conclui-se que a opção de métodos de coloração mais lentos não resultou em execuções de melhor qualidade; no presente trabalho optou-se pela versão MPCG_OI porque dos métodos testados foi um dos mais rápidos apresentando elevados níveis de eficácia.

4.8 Estudo da Complexidade de Execução do Método

Nesta secção pretende-se fazer um estudo da complexidade de execução do método no pior caso e no caso médio.

4.8.1 Pior Caso

No pior caso, para se colorir um vértice terá que se comparar esse vértice com todos os vértices já coloridos. Neste caso a ordem de execução é $O(n^2)$.

Esta situação ocorre quando se pretende colorir um grafo completo, em que são necessárias n cores. Uma vez que a situação do grafo completo não ocorre com frequência, vejamos qual a complexidade de execução no caso médio.

4.8.2 Caso Médio

Suponhamos já coloridos n_ℓ vértices em ℓ classes e seja Δ o grau máximo de entre todos os vértices do grafo. No máximo, para colorir o vértice v fazem-se $d_G(v)$ comparações. No entanto, nem todos os vértices adjacentes a v pertencem ao conjunto dos n_ℓ vértices já coloridos. Assim, no máximo faz-se um número de comparações igual ao número de vértices adjacentes com v já coloridos. Seja X a v.a. que representa o número de vértices adjacentes com v, já coloridos. A distribuição que a v.a. segue é a hipergeométrica.

$$f(x) = P(X = x) = \frac{\begin{pmatrix} Mp \\ x \end{pmatrix} \begin{pmatrix} Mq \\ N - x \end{pmatrix}}{\begin{pmatrix} M \\ N \end{pmatrix}}$$

Com

$$\begin{cases} M = n \\ Mp = d_G(v) \Rightarrow np = d_G(v) \Leftrightarrow p = \dfrac{d_G(v)}{n} \\ Mq = n - d_G(v) \\ N = n_\ell \end{cases}$$

$$f(x) = P(X = x) = \frac{\begin{pmatrix} d_G(v) \\ x \end{pmatrix} \begin{pmatrix} n - d_G(v) \\ n_\ell - x \end{pmatrix}}{\begin{pmatrix} n \\ n_\ell \end{pmatrix}}$$

Assim

$$E(X) = Np = n_\ell \frac{d_G(v)}{n}$$

Assim, quando se colore v (já coloridos n_ℓ vértices) faço em média $\frac{n_\ell}{n} d_G(v)$ comparações. Quando coloro o $i^{\text{ésimo}}$ vértice, então já colori $i-1$ vértices. Assim, considerando as v.a. X_i, tem-se

$$E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \frac{i-1}{n} d_G(v_i) = \frac{1}{n}\sum_{i=1}^{n} i d_G(v_i) - \frac{1}{n}\sum_{i=1}^{n} d_G(v_i) = \frac{1}{n}\sum_{i=1}^{n} i d_G(v_i) - \frac{2m}{n}$$

Como,

$$d_G(v_1) + 2d_G(v_2) + \ldots + n d_G(v_n) \leq \Delta + 2\Delta + \ldots + n\Delta = \frac{n}{2}(\Delta + n\Delta) = \frac{n(n+1)}{2}\Delta$$

Verifica-se

$$\sum_{i=1}^{n} \frac{i-1}{n} d_G(v_i) \leq \frac{n+1}{2}\Delta - \frac{2m}{n}$$

Como m é no mínimo igual a zero, verifica-se:

$$E\left(\sum_{i=1}^{n} X_i\right) \leq \frac{n+1}{2}\Delta$$

Proposição 4.1 *Se Δ é o grau máximo de entre os n vértices e se $T(n)$ for o número de operações elementares esperadas para colorir n vértices, então*

$$T(n) \leq \frac{n+1}{2}\Delta = O(n\Delta)$$

4.9 Conclusão

Usando alguns conceitos da teoria de grafos, foi definido um método de classificação não hierárquica baseado na coloração de grafos, com ajuste que reduz o número de cores, com o objectivo de que as classes de cor encontradas sejam também as classes homogéneas do conjunto de dados. O método de classificação foi associado a um índice de classificabilidade, também definido no contexto de grafos, e em conjunto, pretendem identificar o número e o conteúdo das classes do conjunto de dados. A eficácia do método foi constatada em conjuntos de dados artificiais, evidenciando a sua robustez no que diz respeito a conjuntos de dados com pontos isolados, com classes de formato não esférico e sem uma estrutura em classes. O método mostrou-se estável quando aplicado a várias modificações de um mesmo conjunto de dado. A versão não ajustada apresenta uma ordem de execução de $O(\Delta n)$ e a ajustada apresenta uma ordem de execução de $O(n^2)$.

Capítulo 5

Índices de Avaliação de Classes e Partições

O método proposto permite gerar e avaliar partições, sem fazer qualquer análise de cada uma das classes individualmente. Numa mesma classificação desconhece-se o contributo de cada uma das classes para a identificação da estrutura nos dados, sendo esta informação importante para se ter um conhecimento mais profundo sobre o modo como os elementos estão relacionados entre si. Usando ainda conceitos de teoria de grafos, são apresentados neste capítulo índices de avaliação do isolamento e da densidade de classes, obtidas pelo método de classificação não hierárquica apresentado no capítulo 4, estendido a conjuntos de dados caracterizados por atributos (ou variáveis) mistos. É ainda apresentado um índice de avaliação de partições, que permite escolher a partição que melhor reflecte a estrutura existente nos dados, de entre as candidatas obtidas pelo método de classificação. Consideram-se partições candidatas aquelas para as quais o índice $\tilde{IC}$ apresenta valores indicativos do número de classes, nomeadamente quando existem várias partições com o mesmo $\tilde{IC}$ associado (por exemplo, suponhamos que o método identificou várias partições com $\tilde{IC} = 0$, qual escolher?). Os índices propostos são aplicados a conjuntos de dados artificiais de atributos numéricos e mistos, e comparados com alguns existentes na literatura.

Um dos índices de isolamento apresentados permite definir uma medida de dissemelhança entre classes, que proporciona o desenvolvimento de um ajuste hierárquico baseado em grafos ponderados pela medida de dissemelhança, a usar em alternativa ao ajuste baseado em densidade definido no capítulo anterior. São apresentados alguns exemplos que constatam a eficiência desta nova abordagem.

Comecemos por apresentar alguns índices existentes na literatura.

5.1 Índices de Validação de Classes

Existem na literatura muitas definições de classe, não sendo nenhuma delas universal. Em geral, cada método de classificação pretende encontrar as classes seguindo uma definição particular. No caso geral, as classes de uma "boa" classificação devem ser homogéneas, em que os elementos de uma mesma classe deverão ser similares; e devem ser bem separadas, i.e., entidades dissimilares deverão pertencer a classes diferentes [51].

Uma possível definição de classe é considerá-la como um subconjunto do conjunto de dados, tal que qualquer dissemelhança entre elementos da classe é menor que qualquer dissemelhança entre elementos da classe e elementos fora da classe [122]. Uma outra definição consiste em identificar uma classe como sendo um subconjunto do conjunto de dados, para o qual o diâmetro da classe (máximo das distâncias entre pares de elementos da classe), é menor que o mínimo das distâncias entre os elementos da classes e elementos externos à classe [218]. Outra possível definição é dada em [148] no contexto da teoria de grafos, onde é definido um *(ℓ,r)-cluster*, para algum r, como tendo a propriedade de que numa vizinhança de cada um dos elementos da classe de raio r, têm que existir pelo menos ℓ outros elementos da mesma classe; e todos os elementos do conjunto de dados podem ser ligados por um caminho de comprimento menor ou igual a r. Existem ainda outras definições de classes baseadas em conceitos de teoria de grafos. Por exemplo, uma classe pode corresponder a um clique ([50], [105] e [16]), a um subgrafo conexo ([50], [216] e [1]) ou a um subgrafo conexo maximal ([106]) do grafo definido no conjunto de dados. Muitas outras definições de classes podem ser encontradas na literatura ([146] e [115]).

Cada índice de validação pretende validar uma classe de acordo com a sua própria definição. Em geral, a validação de uma classe engloba o teste de três características principais: densidade, isolamento e estabilidade [168].

5.1.1 Medidas de Heterogeneidade e de Isolamento de Classes numa Partição

Em [91] são apresentados alguns critérios de heterogeneidade e de isolamento de partições, definidos a partir dos mesmos critérios aplicados a cada uma das classes da partição. Assim, se $P = \{C_1, C_2, \ldots, C_k\}$ for uma partição em k classes, então a fórmula geral para os critérios de heterogeneidade são dados por:

$$H(P) = \sum_{r=1}^{k} H(C_r) \quad \text{ou} \quad H(P) = max_{\{r=1,\ldots,k\}}\{H(C_r)\} \tag{5.1}$$

O critério de heterogeneidade da classe C_r, pode ser definido como o quadrado da distância euclideana entre os elementos da classe e o seu centro; como a distância *city-block* entre os elementos da classe e o vector mediana; como a dissemelhança máxima entre pares de elementos da classe; ou ainda como o mínimo, de entre todos os elementos da classe, das somas das distâncias entre cada elemento e todos os outros elementos da classe.

A fórmula geral dos critérios de isolamento é dada por:

$$I(P) = \sum_{r=1}^{k} I(C_r) \quad \text{ou} \quad I(P) = min_{\{r=1,\ldots,k\}}\{I(C_r)\} \tag{5.2}$$

O critério de isolamento da classe C_r, pode ser definido como o mínimo das distâncias entre os elementos da classe e fora da classe, ou como a soma de todos as distâncias entre todos os elementos da classe e todos os elementos pertencentes a outras classes.

Assim, as medidas de isolamento e heterogeneidade de cada classe de uma partição, podem ser combinadas para proporcionarem medidas de adequação de uma partição. As partições óptimas são aquelas que minimizam as medidas de heterogeneidade e maximizam as medidas de isolamento, ou são baseadas numa combinação daquelas medidas.

5.1.2 Índice de Isolamento de *Ling*

No contexto da classificação hierárquica, seja C um *(ℓ,r)-cluster*, sendo r o nível no qual a classe é formada, e r' o nível no qual a classe C é absorvida por outra classe. O *índice de isolamento* de C ([148]), é dado por $r - r'$ e representa o "tempo" que decorre entre a classe ser formada e ser absorvida por outra classe.

Assim, uma classe é compacta se for formada relativamente cedo (quando r é pequeno) para o seu tamanho, sendo isolada se o seu índice de isolamento for grande. O índice de *Ling* baseia-se na ideia de que a duração de vida de uma classe válida, deverá ser superior à duração de vida de uma classe encontrada ao acaso. Tendo como plataforma probabilística um modelo nulo de matrizes aleatórias, prova-se ([148]) que o índice de isolamento segue uma distribuição aproximadamente Geométrica, com parâmetro $p = 1 - \dfrac{n_c(n-n_c)}{N-r}$ onde $N = \left(\begin{array}{c} n \\ 2 \end{array} \right)$, para $(1,r) - clusters$ com n_c elementos.

5.1.3 Índices de Densidade e Isolamento de *Bailey & Dubes*

Considere-se o grafo $G(n,m)$ de n vértices, correspondentes aos n elementos de Ω, e m arcos. Um par de vértices i e j estão ligados por um arco se a dissemelhança entre i e j está entre as m dissemelhanças menores, de entre todos os pares de elementos de Ω. Seja C um subconjunto de Ω pré-definido com r elementos. O objectivo é determinar se C é uma boa classe. Seja E_m o conjunto dos arcos $\{i,j\}$ do grafo $G(n,m)$. A densidade da classe C é definida por [8]: $D_C(m) = \text{card}\{\{i,j\} | i,j \in C, \{i,j\} \in E_m\}$ e representa o número de arcos internos à classe C no grafo $G(n,m)$. Grandes valores de $D_C(m)$ indicam que a classe C é compacta. O isolamento da classe C é definido por [8]: $I_C(m) = \text{card}\{\{i,j\} | i \in C, j \notin C, \{i,j\} \in E_m\}$. Pequenos valores de $I_C(m)$ indicam que é isolada.

Seja I a variável aleatória que representa o número de arcos que ligam os r elementos de uma classe aos outros vértices do grafo e seja D a variável aleatória que representa o número de arcos internos à classe de r elementos. Para uma classe C de tamanho r que não seja o resultado de um algoritmo de classificação, as distribuições destes índices sob a hipótese do grafo aleatório (H_0) são hipergeométricas [91]: $H\left(n_2, m; \frac{r_2}{n_2}\right)$ para o índice de densidade e $H\left(n_2, m; \frac{r(n-r)}{n_2}\right)$ para o índice de isolamento, onde $n_2 \equiv \frac{n(n-1)}{2}$ e $r_2 \equiv \frac{r(r-1)}{2}$.

No entanto, se C foi obtida como o resultado de um algoritmo de classificação, estas distribuições terão de ser modificadas. Neste caso, são sugeridos alguns índices cujo princípio base consiste na comparação dos valores do isolamento e da densidade de uma classe, resultado da aplicação de um algoritmo de classificação, com os valores de uma classe escolhida ao acaso. As fórmulas para os índices de isolamento e densidade são ([91] e [168]):

$$\tilde{I}_C(m) = min\left\{ \binom{n}{r} P(I \leq i_C \mid H_0), 1 \right\} \tag{5.3}$$

$$\tilde{D}_C(m) = min\left\{ \binom{n}{r} P(D \geq d_C \mid H_0), 1 \right\} \tag{5.4}$$

considerando as distribuições hipergeométricas indicadas.

5.1.4 Métodos para Testar a Estabilidade de Classes

Em [168] é proposto um método de validação de uma única classe, através de novos índices de validação que se baseiam na avaliação da estabilidade de uma classe por métodos de reamostragem, consistindo na medição da influência da remoção de certos elementos do conjunto de dados, sobre essa classe. O princípio adoptado é o seguinte: aplica-se o método de partição sobre um grande número de amostras retiradas do conjunto Ω, de seguida verifica-se se os elementos da classe C são, tanto quanto possível, classificados juntos nas diferentes amostras, não se misturando com elementos de outras classes. Este tipo de comportamento é evidente numa classe idealmente estável, tanto do ponto de vista de isolamento como do ponto de vista da densidade.

5.1.5 Estatística de Validação U

Uma outra estatística de validação de uma classe é inicialmente proposta em [156] e frequentemente referida em trabalhos posteriores ([89], [91] e [168]):

$$U_{ijkl} = \begin{cases} 0 & \text{if} \quad d_{ij} < d_{kl} \\ \frac{1}{2} & \text{if} \quad d_{ij} = d_{kl} \\ 1 & \text{if} \quad d_{ij} > d_{kl} \end{cases}$$

Para uma classe C de tamanho r, seja:

$W \equiv \{\{i,j\}|i,j \in C, i<j\}$ o conjunto de $\frac{r(r-1)}{2}$ pares de elementos da classe, e

$B \equiv \{\{k,l\}|k \in C, l \notin C\}$ o conjunto de $r(n-r)$ pares de elementos entre classes.

O índice U global, U_G é definido por:

$$U_G \equiv \sum_{(i,j)\in W} \sum_{(k,l)\in B} U_{ijkl} \tag{5.5}$$

O índice U local, U_L é definido por:

$$U_L \equiv \sum_{i\in C} \sum_{j\in C\setminus\{i\}} \sum_{k\notin C} U_{ijik} \tag{5.6}$$

Quanto mais baixos forem os valores dos índices U_G e U_L melhor é a classe. As classes que apresentem os índices nulos são aquelas que verificam as condições de classe ideal, pela qual as dissemelhanças entre elementos pertencentes às mesmas classes, são sempre menores que entre elementos pertencentes a classes diferentes.

5.1.6 Índices de Isolamento, Densidade e Validação de *Bel Mufti*

No trabalho descrito em [168] foram definidos e testados índices de medição do isolamento, da densidade e da validação de uma classe C, gerada por um método de partições, aplicado a um conjunto de dados Ω a classificar, usando um esquema de reamostragem. Designemos por $\xi_1, \xi_2, \ldots, \xi_N$, N amostras de Ω, obtidas respeitando uma cota de $p\%$ sobre cada classe da partição.

A definição de cada um dos índices aplicados a uma classe é feita de modo externo, a partir de uma partição de referência; ou de modo interno a partir de N partições de Ω em k classes, obtidas a partir de N amostras do conjunto de dados.

Para avaliar de modo externo os índices de isolamento e de densidade da classe C, dispomos da partição $Q = \{Q_1, \ldots, Q_k\}$ conhecida *apriori*. Nestas condições, a classe C será considerada isolada se dois elementos que são classificados separadamente na partição $Q_C = \{C, \bar{C}\}$, também o forem na partição Q conhecida. A classe C será considerada compacta se dois elementos que são classificados juntos na classe C, também o são na partição Q. Os *índices de isolamento e densidade externos*, representam-se por $I_e(C)$ e $H_e(C)$ respectivamente, e são definidos para uma classe C como sendo:

$$I_e(C) = 1 - \frac{2}{n(n-1)} \sum_{j=1}^{k} |C \cap Q_j||\bar{C} \cap Q_j| \tag{5.7}$$

e

$$H_e(C) = 1 - \frac{2}{n(n-1)} \sum_{1 \leq j < \ell \leq k} |C \cap Q_j||C \cap Q_\ell| \tag{5.8}$$

Verifica-se $0 \leq I_e(C) \leq 1$ e $0 \leq H_e(C) \leq 1$. Quanto mais próximos os índices estiverem de 1 mais isolada e mais compacta é a classe. Quanto mais próximos os índices estiverem de 0, menos a classe é isolada e compacta, respectivamente.

Para avaliar de modo interno o isolamento e a densidade da classe C, dispomos de N partições $Q^{(1)}, \ldots, Q^{(N)}$ que foram obtidas pela aplicação do método de classificação sobre N amostras $\xi_1, \ldots, \xi_N$ do conjunto Ω. Para cada amostra $\xi_i (i = 1, \ldots, N)$, obtém-se uma partição, em k classes $Q^{(i)} = \{Q_1^{(i)}, \ldots, Q_k^{(i)}\}$. Nestas condições, a classe C será considerada como isolada se dois elementos que são classificados em classes diferentes na partição $\{C, \bar{C}\}$, também o são na partição $Q^{(i)}$ para algum i. A classe C será considerada como compacta se dois elementos que são classificados juntos na classe C, também o são na partição $Q^{(i)}$ para algum i. Os *índices de isolamento e densidade internos*, designam-se por $I(C)$ e $H(C)$ e são definidos para uma classe C da seguinte maneira:

$$I(C) = 1 - \frac{2}{n(n-1)N} \sum_{i=1}^{N} \sum_{j=1}^{k} |C \cap Q_j^{(i)}||\bar{C} \cap Q_j^{(i)}| \tag{5.9}$$

e

$$H(C) = 1 - \frac{2}{n(n-1)N} \sum_{i=1}^{N} \sum_{1 \leq j < \ell \leq k} |C \cap Q_j^{(i)}||C \cap Q_\ell^{(i)}| \tag{5.10}$$

Verifica-se $0 \leq I(C) \leq 1$ e $0 \leq H(C) \leq 1$. Quanto mais próximo o índice estiver de 1 mais isolada e compacta é a classe. Quanto mais próximo o índice estiver de 0, menos isolada e compacta, respectivamente é a classe.

Dizemos que uma classe C é *válida*, se for ao mesmo tempo isolada e compacta, ou seja, se as duas regras que garantem respectivamente o isolamento e a compacticidade forem satisfeitas. Assim os *índices de validação externo e interno*, representados por $V_e(C)$ e $V(C)$ respectivamente, são definidos para uma classe C da seguinte maneira:

$$V_e(C) = 1 - \frac{2}{n(n-1)} \left(\sum_{j=1}^{k} |C \cap Q_j||\bar{C} \cap Q_j| + \sum_{1 \leq j < \ell \leq k} |C \cap Q_j||C \cap Q_\ell| \right) \tag{5.11}$$

e

$$V(C) = 1 - \frac{2}{n(n-1)N} \sum_{i=1}^{N} \left(\sum_{j=1}^{k} |C \cap Q_j^{(i)}||\bar{C} \cap Q_j^{(i)}| + \sum_{1 \leq j < \ell \leq k} |C \cap Q_j^{(i)}||C \cap Q_\ell^{(i)}| \right) \tag{5.12}$$

Verifica-se $0 \leq V_e(C) \leq 1$ e $0 \leq V(C) \leq 1$. Quanto mais próximos os índices estiverem de 1, mais válida é a classe. Quanto mais próximos estiverem de 0 menos válida é a classe.

5.2 Índices de Validação de Partições: τ_a e τ_w de *Guénoche*

Os índices de validação τ_a e τ_w [98] são baseados no princípio de que uma partição é "boa", se grandes dissemelhanças se verificam entre elementos de classes diferentes, e pequenas dissemelhanças se verificam entre elementos da mesma classe. O índice *rácio de acordo entre pares*, designado por τ_a, mede a qualidade de uma partição pela percentagem de pares de objectos, que pertencendo à mesma classe, apresentam pequenas dissemelhanças; e que pertencendo a classes diferentes, apresentam grandes dissemelhanças. Uma partição P em Ω induz uma bipartição no conjunto dos pares

de elementos do conjunto de dados. Seja (L_e, L_i) essa partição, e sejam N_e e N_i o número de elementos em cada parte. A parte L_e contém todos os pares de objectos que pertencem a classes diferentes, e a parte L_i contém todos os pares de objectos pertencentes à mesma classe. Depois de ordenar os valores das dissemelhanças por ordem decrescente, divide-se a sequência em duas partes, a primeira com os N_e maiores valores e a segunda com os N_i menores valores. Esta partição é designada por (G_d, S_d). Uma partição perfeita em Ω, de acordo com o príncipio anterior, tem as duas bipartições idênticas. O índice *rácio de acordo entre pares* conta a percentagem de pares que pertencem em simultâneo a L_e e a G_d ou a L_i e a S_d. Neste caso, valores do índice perto de 1, identificam classificações com melhor qualidade.

O outro índice de validação é denominado *rácio de pesos*, é designado por τ_w, e é calculado pelas somas das dissemelhanças entre elementos dos quatro tipos de pares: $\sum(L_e)$, $\sum(L_i)$, $\sum(G_d)$ and $\sum(S_d)$, como

$$\tau_w = \frac{\sum(L_e)}{\sum(G_d)} \times \frac{\sum(S_d)}{\sum(L_i)} \tag{5.13}$$

Estes dois racios correspondem ao peso das ligações externas, dividido pelo máximo que pode ser conseguido com N_e valores de dissemelhanças; e pelo peso mínimo de N_i ligações, dividido pelo peso das ligações internas. Este índice toma valores menores ou iguais a 1. Um valor próximo do máximo significa que as dissemelhanças entre classes correspondem às maiores dissemelhanças, e que as dissemelhanças dentro de classes correspondem às menores dissemelhanças.

De seguida apresentam-se alguns índices que foram desenvolvidos no âmbito da presente dissertação.

5.3 Índice de Isolamento Absoluto

O *Índice de Isolamento Absoluto* (IIA) pretende determinar o isolamento de cada classe, no contexto da partição na qual se insere.

5.3.1 Definição do Índice de Isolamento Absoluto

O valor do índice apresentado nesta secção está directamente relacionado com o valor do índice de classificabilidade, pretendendo-se que determine o contributo que cada uma das classes tem para o valor deste índice. Existindo uma relação inversa entre o valor do índice de classificabilidade e a qualidade da partição identificada pelo método, MPCG uma classe deverá ser tanto mais isolada quanto menor for o valor do índice de isolamento absoluto correspondente.

Seja C_i uma das k classes da partição.

Definição 5.1 *O valor do* Índice de Isolamento Absoluto IIA*, da classe C_i representa-se por Δ_i^a, e é dado por:*

$$\Delta_i^a = \sum_{v \in C_i} [d_G^{max}(v) - d_G(v)] \tag{5.14}$$

sendo $d_G^{max}(v)$ e $d_G(v)$, o grau máximo e o grau do vértice v no grafo G representativo dos dados, respectivamente. Se n for o número de vértices do grafo estruturado em k classes, $C_1, \ldots, C_k$ com $n_1, \ldots, n_k$ elementos respectivamente, então

$$\Delta_i^a = n_i(n - n_i) - \sum_{v \in C_i} d_G(v) \tag{5.15}$$

Para cada vértice $v \in C_i$, $d_G^{max}(v) - d_G(v)$ é o número de vértices exteriores a C_i aos quais v dista menos do que α.

Como as classes identificadas no conjunto de dados são também classes de cor no contexto da teoria de grafos, verifica-se que numa mesma classe não existem vértices adjacentes. Um vértice de uma classe só poderá ser adjacente no máximo, a todos os outros vértices fora da classe à qual pertence. Deste modo $d_G^{max}(v) = n - n_i$. Como se pretende a soma dos graus máximos para todos os vértices da classe C_i, vem

$$\begin{aligned} \Delta_i^a &= \sum_{v \in C_i} [d_G^{max}(v) - d_G(v)] = \sum_{v \in C_i} d_G^{max}(v) - \sum_{v \in C_i} d_G(v) \\ &= n_i(n - n_i) - \sum_{v \in C_i} d_G(v) \end{aligned} \tag{5.16}$$

Consideremos a definição do índice de classificabilidade definido na equação (4.17) (Secção 4.3). Relacionando o valor do índice de classificabilidade com o valor do índice de isolamento absoluto, resulta:

$$IC = \frac{1}{2}\sum_{i=1}^{k} \Delta_i^a \tag{5.17}$$

Assim, quanto menor for o valor $\Delta^a{}_i$ para a classe C_i, menor é a sua contribuição para o valor IC e como tal, maior é o seu isolamento absoluto no contexto da partição onde se insere.

Para que o número de elementos nas classes não tenha um peso excessivo no valor do índice, define-se:

$$\tilde{\Delta}_i^a = \frac{\Delta^a{}_i}{n_i(n-n_i)} = 1 - \frac{1}{n_i(n-n_i)} \sum_{v \in C_i} d_G(v) \tag{5.18}$$

Verifica-se que $0 \leq \tilde{\Delta}_i^a \leq 1$, tomando o valor zero para classes isoladas. Quanto maior for o valor do índice, menos isolada é a classe.

O índice de isolamento absoluto definido pelos grafos *k-partites* completos, usa na sua determinação um conceito já usado por alguns autores como por exemplo, *Bailey & Dubes* [8] (Secção 5.1.3). Um dos índices definidos por estes autores é baseado no número de arcos que unem elementos da classe C_i a elementos fora da classe, de um grafo definido nos dados. No grafo definido pelos autores, dois vértices são adjacentes se a dissemelhanças entre eles for menor que um dado parâmetro; considere-se o valor desse parâmetro igual ao valor de α. Valores baixos deste índice indicam classes isoladas. O índice de isolamento absoluto conta o número de arcos em falta para que o grafo seja *k-partite* completo. Tendo em conta que na abordagem utilizada pelo método, dois vértices v_i, v_j são adjacentes se $dist(v_i, v_j) \geq \alpha$, então o índice conta o número de arcos cujas dissemelhanças entre os respectivos vértices não é maior ou igual a α, ou seja, conta o número de arcos com dissemelhança entre vértices menor que α. Assim, com definições complementares para os grafos representativos nos dados, o índice de isolamento absoluto e o índice de *Bailey & Dubes* baseiam-se numa mesma quantidade.

Tabela 5.1: Valores do índice IIA para o conjunto representado na figura 5.1.

	IIA
C_1	0
C_2	0.004663
C_3	0.005371
C_4	0.020833

5.3.2 Aplicação do Índice de Isolamento Absoluto

O índice de isolamento absoluto foi aplicado às classes das partições obtidas pela aplicação do método, a dois conjuntos de dados. A figura 5.1 representa a distribuição em quatro classes de um conjunto de dados obtida pelo método para $\alpha = 97.33817$ e $\tilde{IC} = 0.004261$. A tabela 5.1 mostra os valores do índice de isolamento absoluto para cada uma das classes, indicando que o índice identificou correctamente a classe mais isolada, como sendo a classe C_1.

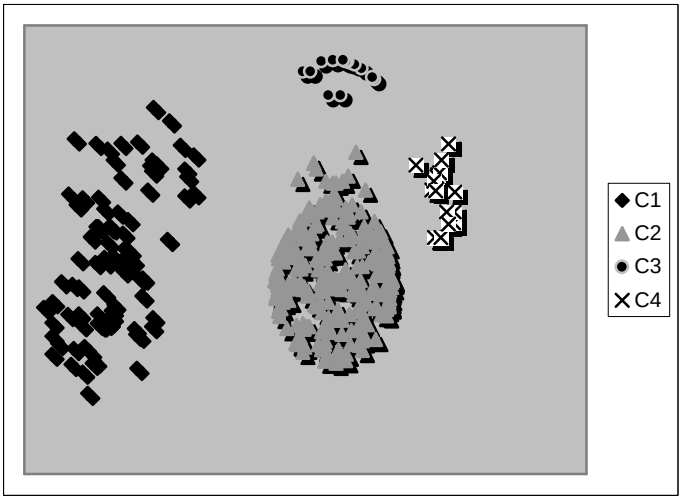

Figura 5.1: Partição em quatro classes obtida pelo método.

A figura 5.3 representa o resultado da aplicação do método ao conjunto de dados representado na figura 5.2. Como se pode ver pela parte superior do gráfico, a única partição que é possível identificar é a partição em duas classes. Assim, repetiu-se a aplicação do método para $\ell_{min} = 16.67248$, $\ell_{max} = 58.46575$ e $H = 50$. O gráfico da parte inferior da figura mostra o resultado obtido. Para $\alpha = 26.702861$ obteve-se a partição em cinco classes com $\tilde{IC} = 0.000999$ e representada na figura 5.2. A tabela 5.2 mostra os resultados obtidos da aplicação do índice de isolamento absoluto para

Tabela 5.2: Valores do índice IIA para o conjunto representado na figura 5.2.

	IIA
C_1	0.00003
C_2	0.000122
C_3	0.001369
C_4	0.002757
C_5	0.000865

cada uma das cinco classes. O índice identificou correctamente a classe C_1 como a mais isolada. A classe menos isolada é identificada pelo índice como sendo a C_4, pois é a classe que está mais perto de um maior número de outras classes.

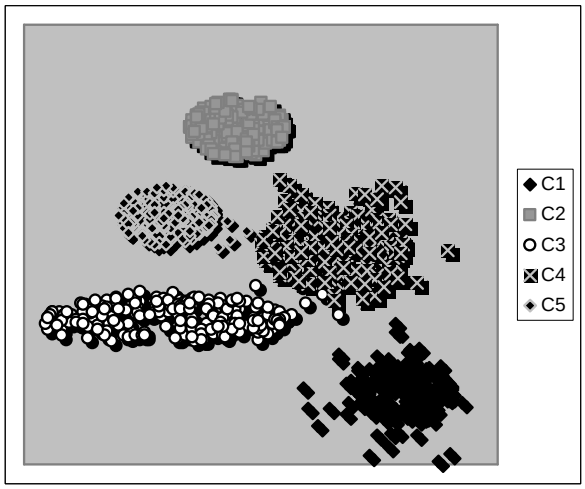

Figura 5.2: Partição em cinco classes obtida pelo método.

5.4 Índice de Isolamento Relativo

O *Índice de Isolamento Relativo* (IIR) pretende determinar o isolamento relativo entre classes de uma partição.

5.4.1 Definição do Índice de Isolamento Relativo

O índice de isolamento relativo, pretende medir o isolamento de uma classe C_i em relação a outra classe da mesma partição. Designemos por $\Delta_i^r(C_j)$ o isolamento da

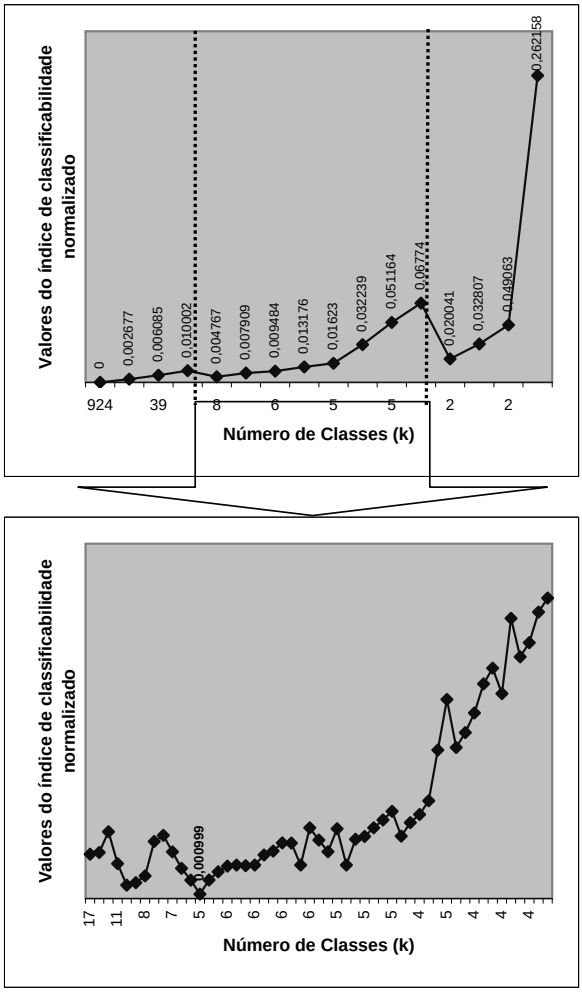

Figura 5.3: Gráfico $\tilde{IC}/k$ obtido para o conjunto de dados representado na figura 5.2.

classe C_i em relação à classe C_j. Num grafo *k-partite*, o grau de um vértice pertencente à classe C_i, consiste na soma do número de arcos que unem o vértice aos elementos de cada uma das classes existentes na partição. Ou seja, o grau de $v \in C_i$ pode ser definido da seguinte maneira:

$$d_G(v) = \sum_{\{j=1, j\neq i\}}^{k} d_G^{C_j}(v) \tag{5.19}$$

sendo $d_G^{C_j}(v)$ o número de arcos que unem o vértice a elementos da classe C_j. Para se obter um grafo *k-partite* completo, cada vértice de $v \in C_i$ deverá ter $d_G^{C_j}(v) = n_j$, para $j = 1, \ldots, k, j \neq i$. Para cada vértice $v \in C_i$, $n_j - d_G^{C_j}(v)$ é o número de vértices de C_j que distam de v menos do que α.

Definição 5.2 *O valor do* Índice de Isolamento Relativo IIR*, da classe C_i em relação à classe C_j representa-se por $\Delta_i^r(C_j)$, é calculado como sendo a diferença entre o grau máximo relativo à classe C_j e o grau de cada um dos vértices de C_i:*

$$\Delta_i^r(C_j) = \sum_{v \in C_i} [n_j - d_G^{C_j}(v)] \tag{5.20}$$

sendo $d_G^{C_j}(v)$ o grau do vértice v relativamente a C_j. Se n for o número de vértices do grafo estruturado em k classes, $C_1, \ldots, C_k$ com $n_1, \ldots, n_k$ elementos respectivamente, então

$$\Delta_i^r(C_j) = n_i n_j - \sum_{v \in C_i} d_G^{C_j}(v) \tag{5.21}$$

Normalizando, vem:

$$\tilde{\Delta}_i^r(C_j) = 1 - \frac{1}{n_i n_j} \sum_{v \in C_i} d_G^{C_j}(v) \tag{5.22}$$

Quanto maior for o valor do índice, menor é o isolamento relativo entre as classes.

Finalmente, refira-se que a medida de isolamento entre classes definida nesta secção, depende do valor do parâmetro de controle α, verificando-se a seguinte relação:

$$\Delta_i^r(C_j) = 0 \;\Leftrightarrow\; \forall v \in C_i, \;\; dist(v,u) \geq \alpha, \forall u \in C_j \tag{5.23}$$

Neste caso, as classes C_i e C_j estão afastadas, para o valor de α.

5.4.2 Aplicação do Índice de Isolamento Relativo

O índice foi aplicado a vários conjuntos de dados indicando um bom desempenho na determinação do isolamento entre classes. A tabela 5.3 indica o isolamento entre todas as classes representadas na figura 5.4. Uma classe isolada de todas as outras do conjunto de dados, apresenta uma linha só de zeros como é o caso a classe C_2. As classes C_1 e C_5 e as classes C_3 e C_4 não estão isoladas, verificando-se ainda que as classes C_3 e C_4 estão mais próximas do que as classes C_1 e C_5, como se pode constatar

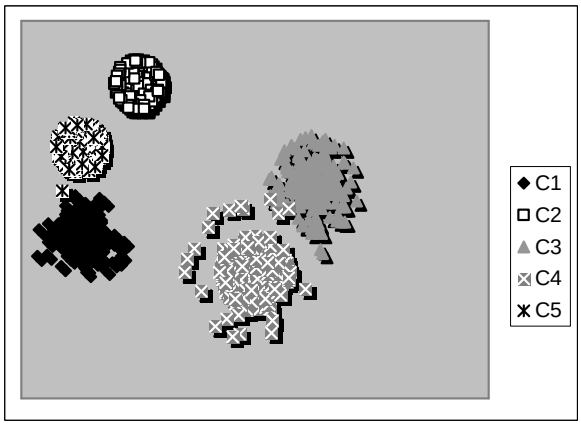

Figura 5.4: Partição em cinco classes obtida pelo método.

Tabela 5.3: Valores dos índices IIR para o conjunto representado na figura 5.4.

	C_1	C_2	C_3	C_4	C_5
C_1	0	0	0	0	0.002435
C_2	0	0	0	0	0
C_3	0	0	0	0.004365	0
C_4	0	0	0.004365	0	0
C_5	0.002435	0	0	0	0

pela figura.

A tabela 5.4 indica o isolamento entre todas as classes representadas na figura 5.2. As classes C_3 e C_4 são as menos isoladas, sendo o par de classes C_4 e C_5 o segundo par menos isolado. A classe C_4 é a menos isolada de um maior número de outras classes. As classes C_2 e C_5 só não são isoladas de uma outra classe (C_4).

Tabela 5.4: Valores dos índices IIR para o conjunto representado na figura 5.2.

	C_1	C_2	C_3	C_4	C_5
C_1	0	0	0.000022	0.00011	0
C_2	0	0	0	0.000563	0
C_3	0.000022	0	0	0.006078	0
C_4	0.00011	0.000563	0.006078	0	0.00393
C_5	0	0	0	0.00393	0

5.5 Índice de Densidade

O *Índice de densidade* (ID) mede a densidade ou compacidade de uma classe. Esta medida está relacionada com o facto de que numa vizinhança de cada elemento, a percentagem de elementos da classe à qual o elemento pertence, deverá ser maior do que a percentagem de elementos de outras classes. Esta medida pretende favorecer a identificação de classes densas não necessariamente esféricas.

5.5.1 Definição do Índice de Densidade

Considere-se uma partição em k classes obtida para o valor do parâmetro de controle α. Seja

$$\frac{|B(v,\tilde{\alpha}_i) \cap C_i|}{n_i - 1} \; ; \; v \in C_i \tag{5.24}$$

a percentagem de elementos de C_i, diferentes de v, que pertencem à bola $B(v,\tilde{\alpha}_i)$, centrada em v e de raio $\tilde{\alpha}_i$. O valor $\tilde{\alpha}$ é o parâmetro de controlo do índice de densidade que, se nenhum outro valor lhe for atribuído, toma o mesmo valor de α. Pretende-se calcular esta quantidade para todos os elementos de C_i. Sendo assim, a densidade de uma classe C_i é dada por:

$$ID'_i := \sum_{v \in C_i} \frac{|B(v,\tilde{\alpha}_i) \cap C_i|}{n_i(n_i - 1)} \tag{5.25}$$

onde n_i é o número de elementos da classe C_i. Tem-se:

$$\frac{1}{n_i - 1} \leq ID'_i \leq \frac{n_i}{n_i - 1} \tag{5.26}$$

Normalizando vem

$$\tilde{ID}_i = ID'_i - \frac{1}{n_i - 1} \tag{5.27}$$

e então

$$0 \leq \tilde{ID}_i \leq 1 \tag{5.28}$$

tomando o valor 1 se todas as bolas contiverem todos os elementos da classe. Isto é, quanto mais próximo o valor de $\tilde{ID}_i$ estiver de 1 mais densa é a classe.

O número de elementos da classe tem muita influência no valor do índice de densidade, levando a que em geral, classes com maior número de elementos sejam consideradas as mais densas. Para diminuir essa tendência dividiu-se o valor de índice por $n - n_i$ e assim o valor do índice é definido por:

$$ID_i := \frac{\tilde{ID}_i}{n - n_i} \tag{5.29}$$

ou seja,

$$ID_i = \sum_{v \in C_i} \frac{|B(v, \tilde{\alpha}) \cap C_i|}{n_i(n_i - 1)(n - n_i)} - \frac{1}{(n_i - 1)(n - n_i)} \tag{5.30}$$

5.5.2 Parâmetro de Controlo do Índice de Densidade

Definamos um outro índice, denominado *limite de conexão.*

Definição 5.3 *O* limite de conexão *de uma classe C_i, designa-se por LC_i (designado em [98] por s_k) e é dado pela distância máxima entre os vizinhos mais próximos, ou seja,*

$$LC_i = max_{\{v \in C_i\}}\{min_{\{u \in C_i\}}\{dist(v, u)\}\}$$

O parâmetro de controlo $\tilde{\alpha}$ de densidade, pode assumir os valores:

- $\tilde{\alpha}_i = \alpha$ ou
- $\tilde{\alpha}_i = LC_i$ ou
- $\tilde{\alpha}_i = max\{\alpha, LC_i\}$ ou
- $\tilde{\alpha}_i = min\{\alpha, LC_i\}$

Tabela 5.5: Valores dos índices W e ID para o conjunto representado na figura 5.5.

	W	ID
C_1	0.45695	0.455313
C_2	1	0.994845
C_3	1	0.994792

para $i = 1, \ldots, k$.

Os resultados obtidos da aplicação do índice a vários conjuntos de dados, para as diferentes opções do seu parâmetro de controlo de densidade, mostraram que o desempenho do índice não varia significativamente para as diferentes opções possíveis. Assim, em todos os exemplos apresentados considerou-se $\tilde{\alpha}_i = \alpha$ para $i = 1, \ldots, k$.

5.5.3 Aplicação do Índice de Densidade

O índice de densidade foi comparado com o índice de densidade de *Bailey & Dubes* [8] (Secção 5.1.3), designado por W e também baseado em teoria de grafos. Neste caso, existe um arco entre dois vértices se a dissemelhança entre eles for menor ou igual a um dado parâmetro. Este índice conta o número de arcos que unem elementos de uma mesma classe C e assume valores altos para classes compactas ou densas.

Os dois índices foram testados em alguns conjuntos de dados. Os resultados mostraram que o índice de densidade ID é mais eficaz na identificação de pequenas diferenças de densidade entre classes. Por exemplo, a figura 5.5 apresenta um conjunto de dados em que as classes mais densas são as que têm menor número de elementos (classes C_2 e C_3) e a menos densa é a que tem maior número de elementos. Como se pode ver na tabela 5.5, o índice W identificou correctamente a classe menos densa, não sendo capaz de diferenciar as outras duas classes. O índice ID identificou correctamente a classe mais densa como sendo a classe C_2.

5.6 Índice de Avaliação de Partições

O *Índice de Avaliação de Partições (IAP)* pretende avaliar uma partição baseado no pressuposto de que o número de vizinhos comuns entre elementos de uma mesma

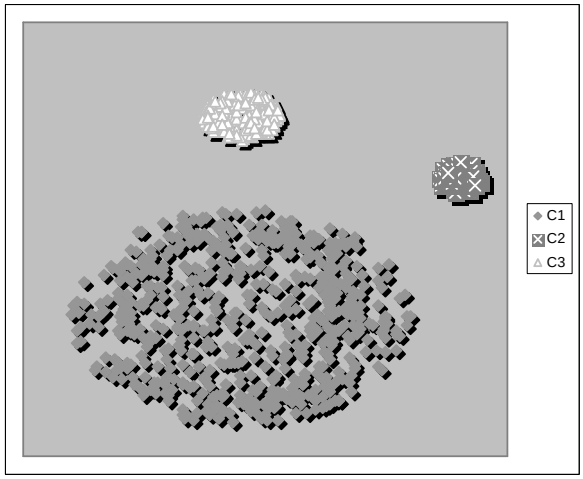

Figura 5.5: Partição em três classes obtida pelo método.

classe, deverá ser maior do que o número de vizinhos comuns entre elementos de classes diferentes.

5.6.1 Definição do Índice de Avaliação de Partições

A *Matriz de Adjacências* $A = [a_{ij}]$, de um grafo G, é uma matriz $n \times n$ binária, simétrica, com:

$$a_{ij} = \begin{cases} 1 & \text{se} \quad \{v_i, v_j\} \in E(G) \\ \\ 0 & \text{se} \quad \{v_i, v_j\} \notin E(G) \end{cases} \tag{5.31}$$

Assim $a_{i.} = \sum_{j=1}^{n} a_{ij} = |Viz(v_i)|$ representa o número de vizinhos de v_i, ou seja, o número de vértices v_j tais que $\{v_i, v_j\} \in E(G)$. Tem-se ainda $a_{i.} = d_G(v_i)$, o grau do vértice v_i.

Seja a_i a coluna i de A e a_j a coluna j de A, a quantidade

$$a_i \cdot a_j = |Viz(v_i) \cap Viz(v_j)| \tag{5.32}$$

representa o *número de vizinhos comuns* entre v_i e v_j.

Sejam $u, v \in C_i$ e $|C_i| = n_i$. O número máximo de vizinhos comuns de u e de v é $n - n_i$, uma vez que estes elementos só podem estar ligados a elementos não pertencentes à classe C_i. Se somarmos o número máximo de vizinhos entre todos os pares de elementos da classe C_i temos:

$$\begin{pmatrix} n_i \\ 2 \end{pmatrix} (n - n_i) \quad = \quad \frac{n_i(n_i - 1)(n - n_i)}{2} \tag{5.33}$$

Assim, a parcela do índice de avaliação de partições correspondente aos vizinhos comuns dentro das classes, representa-se por IAP^d e é dada pela soma do número de vizinhos comuns entre todos os elementos de uma mesma classe, normalizado pelo número máximo de vizinhos comuns. Para cada uma das classes faz-se uma reordenação da matriz de adjacências, de modo a que os n_i elementos da classe C_i estejam nas n_i primeiras linhas da matriz. Assim, o valor da parcela IAP^d é dado por:

$$IAP^d = \frac{2}{k} \sum_{i=1}^{k} \left\{ \frac{\sum_{j=1}^{n_i - 1} \sum_{\ell=j+1}^{n_i} a_j \cdot a_\ell}{n_i(n_i - 1)(n - n_i)} \right\} \tag{5.34}$$

Verifica-se $0 \leq IAP^d \leq 1$, tomando o valor 1 na situação ideal, na qual todos os pares de elementos de cada uma das classes tem o número máximo de vizinhos comuns.

Sejam $u \in C_i$ e $v \in C_j$ com $|C_i| = n_i$ e $|C_j| = n_j$, $i \neq j$. O número máximo de vizinhos comuns de u e de v é $n - n_i - n_j$, uma vez que estes elementos só podem estar simultaneamente ligados a elementos não pertencentes à classe C_i nem à classe C_j. Se somarmos o número máximo de vizinhos entre todos os pares de elementos da classe C_i e C_j temos:

$$(n - n_i - n_j) n_i n_j \tag{5.35}$$

Para cada par de classes C_i e C_j, faz-se uma reordenação da matriz de adjacências, de modo a que os n_i elementos da classe C_i estejam nas n_i primeiras linhas da matriz, estando os n_j elementos da classe C_j nas n_j linhas seguintes. Assim, a parcela do

índice de avaliação de partições correspondente a vizinhos comuns entre elementos pertencentes a classes diferentes, representa-se por IAP^e e é dada por:

$$IAP^e = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \left\{ \frac{\sum_{p=1}^{n_i} \sum_{q=n_i+p}^{n_i+n_j} a_p \cdot a_q}{(n-n_i-n_j)n_i n_j} \right\} \tag{5.36}$$

Verifica-se $0 \leq IAP^e \leq 1$, pretendendo-se que uma partição é tanto melhor, quanto maior for a diferença entre IAP^d e IAP^e.

O *Índice de Avaliação de Partições (IAP)* define-se por

$$IAP = IAP^d - IAP^e \tag{5.37}$$

tomando valores $-1 \leq IAP \leq 1$. Se $0 \leq IAP \leq 1$, então IAP^d é maior ou igual a IAP^e e corresponde a uma boa partição, tendo preferencialmente valores perto de 1. Caso contrário, a partição não deverá ser considerada válida.

5.7 Aplicação dos Índices

Consideremos o conjunto de 500 elementos e dois atributos numéricos representado na figura 5.8. O gráfico 5.6 representa o resultado da aplicação do método a este conjunto de dados. Como se pode verificar pela análise do gráfico, foram identificadas três partições para as quais se verifica $\tilde{IC} = 0$, ou seja, três partições candidatas a representar o conjunto de dados, indicadas na tabela 5.6. Para decidir sobre qual das três partições candidatas é a preferida, aplicaram-se os índices de avaliação de classes e partições definidas neste capítulo, assim como, alguns outros presentes na literatura. Usou-se o índice τ_a definido em [98] (Secção 5.2) que mede a qualidade de uma partição, calculando a percentagem de pares de elementos, que pertencendo à mesma classe, distam menos do que um dado parâmetro e que pertencendo a classes diferentes, distam mais do que o mesmo parâmetro. Neste caso, o índice apresenta valores próximos de um para partições de boa qualidade.

Tabela 5.6: Três possíveis partições identificadas pelo método.

	α	k	$\tilde{IC}$
P_1	14.13847	5	0
P_2	66.69234	4	0
P_3	125.8154	2	0

Tabela 5.7: Avaliação das partições obtidas pelo método.

	P_1	P_2	P_3
τ_a	1	0.994	0.852
IAP	1	0.9964	0.9958

Como se pode ver na tabela 5.7, quer o índice τ_a, quer o índice IAP identificaram como a melhor partição a de cinco classes.

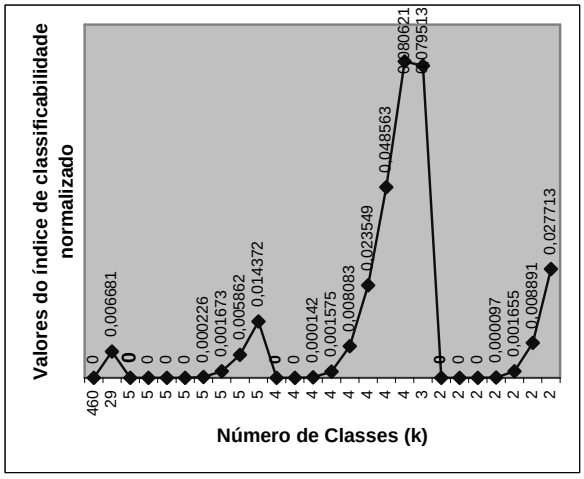

Figura 5.6: Gráfico $\tilde{IC}/k$ obtido para o conjunto de dados representado na figura 5.8.

A este conjunto de dados foram introduzidos dois atributos categóricos, seguindo a mesma distribuição de probabilidades em elementos das mesmas classes, e distribuições diferentes entre elementos de classes diferentes. Pretende-se deste modo, que a introdução dos dois novos atributos, não modifique significativamente a distribuição dos elementos nas classes. A distribuição de probabilidades está representada na tabela 5.8. Aplicou-se o método com a função dissemelhança definida em [111] (Secção 3.9, Equação (3.20)) para atributos mistos. O gráfico resultante da aplicação do método é apresentado na figura 5.7. Como indica a tabela 5.9, foram identificadas quatro partições que, segundo o método utilizado, podem representar o conjunto de dados a

Tabela 5.8: Distribuição de probabilidades das categorias dos atributos do exemplo com dois atributos numéricos e dois categóricos, para cada uma das cinco classes.

Atributo 1			
'a'	'b'	'c'	'd'
0.40	0.04	0.28	0.28
0.10	0.05	0.12	0.73
0.30	0.23	0.17	0.30
0.18	0.26	0.31	0.25
0.06	0.12	0.44	0.38

Atributo 2			
'e'	'f'	'g'	'h'
0.21	0.23	0.18	0.38
0.09	0.16	0.32	0.43
0.51	0.10	0.11	0.28
0.15	0.29	0.26	0.30
0.36	0.12	0.28	0.24

estudar.

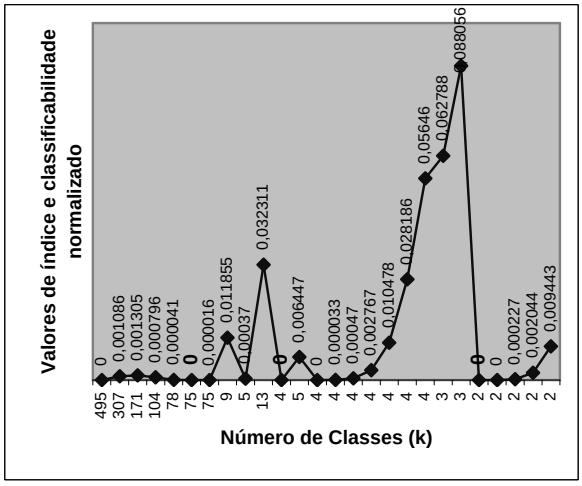

Figura 5.7: Gráfico $\tilde{IC}/k$ obtido para o conjunto de dados representado na figura 5.8, mas com mais dois atributos categóricos.

A tabela 5.10 representa os valores dos índices de avaliação de partições. O índice τ_a identifica a melhor partição a de setenta e cinco classes, enquanto que o IAP identifica como melhor como sendo a de quatro classes. Tendo em conta a distribuição do conjunto representado na figura 5.8, que corresponde à distribuição dos elementos considerando somente os atributos numéricos, e tendo em consideração que o aumento de dimensão introduzindo dois novos atributos categóricos, em princípio, não modificou significativamente a estrutura; existe a tendência de considerar como melhores partições as de quatro e cinco classes. Estas são aliás as duas partições para as quais o índice IAP apresentou valores mais elevados. Vejamos se os índices de avaliação de classes nos fornecem alguma informação, que justifique a escolha da partição em

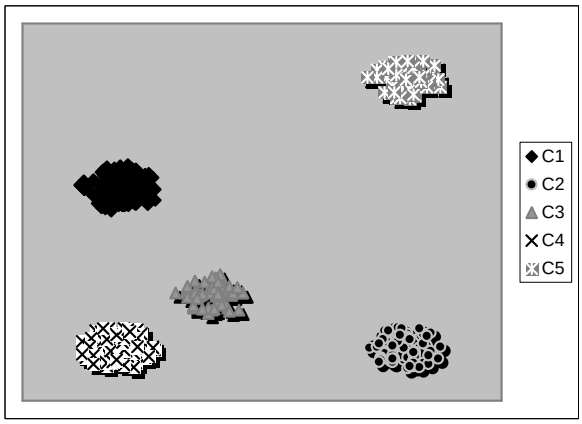

Figura 5.8: Partição em cinco classes obtida pelo método.

Tabela 5.9: Quatro possíveis partições identificadas pelo método.

	α	k	$\tilde{IC}$
P_1	34.27643	75	0
P_2	54.24229	5	0.00037
P_3	67.55286	4	0
P_4	140.761	2	0

quatro classes (P3), como sendo a preferida em relação à partição em cinco classes (P2). O índice de isolamento absoluto considerou todas as classes da partição em quatro classes isoladas, isto é, $\Delta_i^a = 0$ $(i = 1, 2, 3, 4)$ para todas as classes, assim como o IIR considerou que o isolamento relativo entre todas as classes é máximo, ou seja, $\Delta_i^r(C_j) = 0$ para $i, j = 1, 2, 3, 4$ e $i \neq j$. Quanto à partição em cinco classes a tabela 5.12 mostra que todas as classes são isoladas excepto duas, conclusão confirmada pelo índice de isolamento IIR (tabela 5.11) e pelo índice de isolamento absoluto (tabela 5.12), que indica que duas das cinco classes não estão muito isoladas. Assim, a reunião destas duas classes não muito isoladas favorecer o aparecimento de uma partição com melhores índices de avaliação. No entanto, pelos valores apresentados pelos índices τ_a e IAP, qualquer das partições poderia ser considerada para o conjunto de dados, excepto a partição P_4.

Tabela 5.10: Avaliação das partições obtidas pelo método.

	P_1	P_2	P_3	P_4
τ	1	0.993	0.99	0.851
IAP	0.9832	0.9906	0.9929	0.881

Tabela 5.11: Valores do IIR entre classes da partição com 5 classes.

	C_1	C_2	C_3	C_4	C_5
C_1	0	0	0	0	0
C_2	0	0	0	0	0
C_3	0	0	0	0.006	0
C_4	0	0	0.006	0	0
C_5	0	0	0	0	0

Tabela 5.12: Valores do IIA para as classes da partição em 5 classes.

	IIA
C_1	0
C_2	0
C_3	0.001426
C_4	0.000868
C_5	0

5.8 Ajuste Hierárquico

Recorde-se que o método de partição proposto é constituído por duas fases. Em primeiro lugar é aplicado ao conjunto de dados um algoritmo sequencial obtendo-se k_1 classes de cor, tais que $k_1 \geq k$. Estas classes de cor são em seguida alvo de um processo de ajuste baseado em densidade que pretende reduzir o número de classes de cor, identificando no final as classes homogéneas (Secção 4.2.3).

O índice de isolamento relativo permite definir uma medida de dissemelhança entre classes que serve de base ao desenvolvimento de um método de ajuste hierárquico, proposto em alternativa ao método de ajuste baseado em densidade.

5.8.1 Dissemelhança entre Classes

O índice de isolamento relativo entre classes definido na secção anterior pode ser usado para definir uma dissemelhança entre duas classes.

Definição 5.4 *Sejam C_i e C_j duas classes de n_i e n_j elementos respectivamente,*

define-se $\delta(C_i, C_j)$ *como uma medida de dissemelhança entre classes do seguinte modo:*

$$\delta(C_i, C_j) = \begin{cases} 0 & se \quad i = j \\ 1 - \dfrac{\Delta_i^r(C_j)}{n_i n_j} & se \quad i \neq j \end{cases} \tag{5.38}$$

sendo $\Delta_i^r(C_j)$ *o índice de isolamento da classe* C_i *relativamente à classe* C_j*.*

Usando (5.20) tem-se:

$$\delta(C_i, C_j) = \begin{cases} 0 & \text{se} \quad i = j \\ \dfrac{\sum\limits_{v \in C_i} d_G^{C_j}(v)}{n_i n_j} & \text{se} \quad i \neq j \end{cases} \tag{5.39}$$

Esta medida de dissemelhança tem as seguintes propriedades:

1. $0 \leq \delta(C_i, C_j) \leq 1$, para $i, j = 1, \ldots, k$. Valores perto do mínimo representam classes menos isoladas, valores perto do máximo representam classes mais isoladas (mais dissimilares).
2. $\delta(C_i, C_i) = 0$, para $i = 1, \ldots, k$
3. $\delta(C_i, C_j) = \delta(C_j, C_i)$, para $i, j = 1, \ldots, k$

5.8.2 Formação de Grafos Ponderados

Suponhamos que foi aplicado ao conjunto de dados o algoritmo sequencial obtendo-se k_1 classes de cor tais que $k_1 \geq k$. Usando a função de dissemelhança definida em (5.39), é possível construir um grafo ponderado no conjunto de dados com k_1 classes de cor, em que as k_1 classes constituem os vértices do grafo ponderado e cada arco tem custo igual à dissemelhança entre as classes respectivas. Seja GP o Grafo Ponderado assim definido:

1. $V(GP) = \{v_1, \ldots, v_{k_1}\}$ são os k_1 vértices correspondentes às k_1 classes de cor C_i, obtidas pelo método de coloração sequencial.
2. $E(GP) = \{e_1, \ldots, e_N\}$ são os $N = \binom{k_1}{2}$ arcos de custos δ_{ij} tais que $\delta_{ij} = \delta(C_i, C_j)$

5.8.3 Definição do Ajuste Hierárquico

Partindo do grafo ponderado definido a partir da partição em k_1 classes de cor (em geral $k_1 \geq k$), obtidas pelo algoritmo de coloração sequencial ao conjunto de dados, aplica-se um processo hierárquico no qual em cada passo se reunem as classes de menor dissemelhança, ou seja, menos afastadas.

Esquematicamente o método com ajuste hierárquico funciona da maneira que se segue:

1. Aplica-se ao conjunto de dados Ω o algoritmo de coloração sequencial obtendo-se k_1 classes de cor, sejam $C_1, \ldots, C_{k_1}$
2. Cria-se o grafo ponderado completo (GP) cujos vértices são as k_1 classes obtidas no passo anterior. Existe um arco a unir dois quaisquer vértices (classes), de custo associado igual à dissemelhança entre as classes. Designemos por e_{ij} o arco que une as classes C_i e C_j e por c_{ij} uma função tal que $c_{ij} = c(e_{ij}) = \delta(C_i, C_j)$ representa o custo do arco e_{ij}, ou seja, que a cada arco faz corresponder a dissemelhança entre as classes
3. Aplicação do algoritmo de classificação hierárquica ascendente em que a medida de comparação entre classes é a função δ.

Com esta abordagem obtém-se para cada valor de α não uma partição, como no caso do ajuste de densidade, mas uma hierarquia de partições. Podem considerar-se três opções de execução:

1. *É dado o valor de α e de k.* Neste caso, a partir do grafo ponderado definido para α, o método reúne sucessivamente classes até que se obtenha uma partição em k classes.
2. *O valor de α é variável e é dado o valor de k.* Para cada valor de α ($\ell_{min} \leq \alpha \leq \ell_{max}$), o método reúne sucessivamente classes até que se obtenha uma partição em k classes à qual é associado um valor de $\tilde{IC}$. Com esta opção obtemos H (sendo H o número de valores diferentes de α testados) partições em k classes às quais são associados valores de $\tilde{IC}$ que permitem escolher a melhor partição em k classes.
3. *O valor de α é variável e não é dado o valor de k.* Para cada valor de α, e para $1 \leq k \leq k_1$ o método reúne sucessivamente duas classes de cada vez, estando

associado a cada partição obtida um valor de $\tilde{IC}$. Concretamente, começa-se com uma partição em $k = k_1$ classes de cor à qual é associado um valor de $\tilde{IC}$. No passo seguinte obtemos uma partição em $k_1 - 1$ classes, também com um valor de $\tilde{IC}$ associado. Prossegue-se sucessivamente até $k = 1$. Assim, para cada valor de α, escolhe-se a partição em k classes indicado pelo valor de $\tilde{IC}$. Com esta opção determinam-se $(k_1 - 1) \times H$ partições.

Nos casos (1) e (2), considerando k fixo, o processo de reunião de classes pára se numa etapa a dissemelhança entre todos os pares de classes for máxima.

O ajuste hierárquico permite obter partições em k classes com o valor de k fixo *a priori*. No caso do ajuste de densidade não se sabe à partida, quantas classes vai ter a partição indicada pelo valor de $\tilde{IC}$.

5.8.4 Aplicação do Ajuste Hierárquico

Nesta secção apresentam-se alguns exemplos para os quais o método de ajuste acima descrito apresenta algumas vantagens em relação ao método baseado em densidade.

Primeira opção: α conhecido, k conhecido

O novo método de ajuste foi aplicado a todos os conjuntos de dados testados até ao momento. Para todos os casos, fixos os valores adequados de α e de k obtidos pelo método de ajuste de densidade, obtiveram-se exactamente as mesmas partições e os mesmos valores de $\tilde{IC}$.

Segunda opção: α variável, k conhecido

Recordemos as experiências da Secção 4.5, nas quais foram gerados 100 conjuntos de dados para cada um dos conjuntos de parâmetros associados. Foi aplicado o ajuste hierárquico aos 100 conjuntos de dados de cada um dos exemplos. Cada partição identificada pelo método foi comparada com a partição de referência usando o índice IRC. Com o objectivo de comparar o método com alguns métodos clássicos, esta experiência repetiu-se para o algoritmo das *k-médias* e para o método de classificação hierárquica ascendente com índice da média.

Tabela 5.13: Média e desvio padrão dos valores de IRC para 100 execuções do conjunto de dados com classes de forma alongada (Secção 4.5.1).

	Média	Desvio Padrão
Ajuste de Densidade	0.989	0.038
Ajuste Hierárquico / k = 4	1	0
k-Médias / k = 4	0.959	0.094
Índice da Média / k = 4	0.995	0.117

Tabela 5.14: Média e desvio padrão dos valores de IRC para 100 execuções do conjunto de dados com pontos isolados (Secção 4.5.2).

	Média	Desvio Padrão
Ajuste de Densidade	0.922	0.077
Ajuste Hierárquico / k = 5	0.843	0.228
k-Médias / k = 5	0.775	0.127
Índice da Média / k = 5	0.876	0.070

No caso do exemplo com classes alongadas (figura 4.3) os resultados estão representados na tabela 5.13. Pelos valores apresentados na tabela, conclui-se que para este caso, o método com ajuste hierárquico apresenta o melhor desempenho. No caso de exemplo com pontos isolados (figura 4.5) os resultados estão representados na tabela 5.14. Pelos valores apresentados na tabela, conclui-se que para este caso, o método com ajuste de densidade apresenta o melhor desempenho. No caso de exemplo com classes de forma não convexa (figura 4.7) os resultados estão representados na tabela 5.15. Também para este exemplo o ajuste hierárquico apresentou um melhor desempenho. Assim, para estes exemplos o ajuste hierárquico mostrou-se mais eficaz na detecção

Tabela 5.15: Média e desvio padrão dos valores de IRC para 100 execuções do conjunto de dados com duas classes de formas não convexas (Secção 4.5.3).

	Média	Desvio Padrão
Ajuste de Densidade	0.942	0.089
Ajuste Hierárquico / k = 2	1	0
k-Médias / k = 2	0.525	0.001
Índice da Média / k = 2	0.855	0.117
Índice do Mínimo / k = 2	1	0

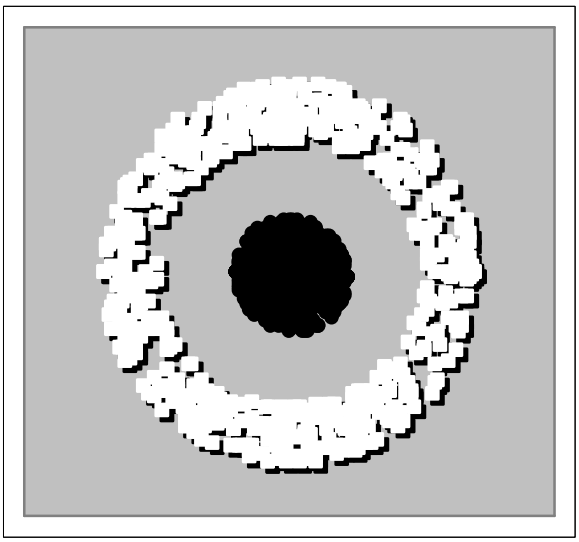

Figura 5.9: Partição em duas classes obtidas pelo método com ajuste hierárquico.

Tabela 5.16: Parâmetros do gerador de dados para as classes do conjunto representado na figura 5.9.

p	a	b	r_1	r_2	r_3	p_1	p_2	p_3
40	100	100	20	20	20	100	0	0
60	100	100	50	70	70	0	100	0

de classes de formas arbitrárias, não sendo, no entanto, tão eficiente como o ajuste de densidade na detecção de classes em conjuntos com pontos isolados.

Considere-se ainda o conjunto de dados com 1000 elementos representados na figura 5.9 gerado com os parâmetros indicados na tabela 5.16. O algoritmo de classificação com ajuste de densidade mostrou-se incapaz de identificar as duas classes correctamente. Como se pode concluir pela análise do gráfico da figura 5.10, não existe um valor que nos permita seleccionar uma partição com alguma segurança. Pelas regras seguidas até ao momento, a partição que melhor descreve os dados é dada para $\alpha = 14.92926$ correspondendo a $\tilde{IC} = 0.005075$ e $k = 12$. Nesta partição o círculo do centro constitui uma classe, sendo a coroa circular dividida em 11 classes. A aplicação do método com ajuste hierárquico para $k = 2$ originou o gráfico representado na figura 5.11. Pela análise deste gráfico identificam-se algumas partições correspondentes ao valor de $\tilde{IC} = 0$. Todas as partições em duas classes para as quais se verificou $\tilde{IC} = 0$ são iguais e identificam correctamente a estrutura do conjunto. Assim, seleccionando um dos valores de α para os quais $\tilde{IC} = 0$, por exemplo $\alpha = 9.357553$, obtém-se

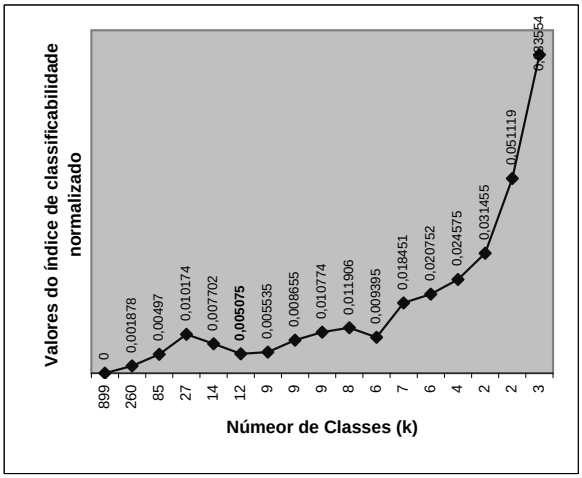

Figura 5.10: Gráfico $\tilde{IC}/k$ obtido para o conjunto de dados representado na figura 5.9 com ajuste de densidade.

a partição representada na figura 5.9. Este exemplo mostra uma maior vantagem do ajuste hierárquico relativamente ao ajuste baseado em densidade. No entanto, recorde-se que na opção da abordagem hierárquica usada neste exemplo, foi indicado *a priori* o número de classes existentes no conjunto de dados.

Terceira opção: α variável, k desconhecido

Nesta opção de execução do ajuste hierárquico, o número de classes do conjunto de dados não é conhecido *a priori*. Neste caso, para cada α obtém-se uma hierarquia de partições. O método selecciona a partição em k classes que minimiza o valor do índice $\tilde{IC}$. Esta abordagem foi aplicada aos conjuntos de dados representados nas figuras 4.3, 4.5, 4.7 e 5.9.

Os resultados estão representados na tabela 5.19. Para cada valor de α, a partição em k classes apresentada corresponde a $\tilde{IC} = 0$. No caso do exemplo das classes alongadas representado na figura 4.3 e cujos resultados estão na primeira e segunda colunas da tabela, opta-se por $k = 4$ uma vez que se obteve esta partição um maior número de vezes. Todas as partições correspondentes a quatro classes são iguais à partição de referência. No caso do conjunto com pontos isolados representado na figura 4.5 e cujos resultados estão na terceira e quarta colunas da tabela, como não existe uma partição que é obtida um maior número e vezes, usa-se o índice IAP para seleccionar a partição representativa da estrutura dos dados. A tabela 5.17 mostra os valores

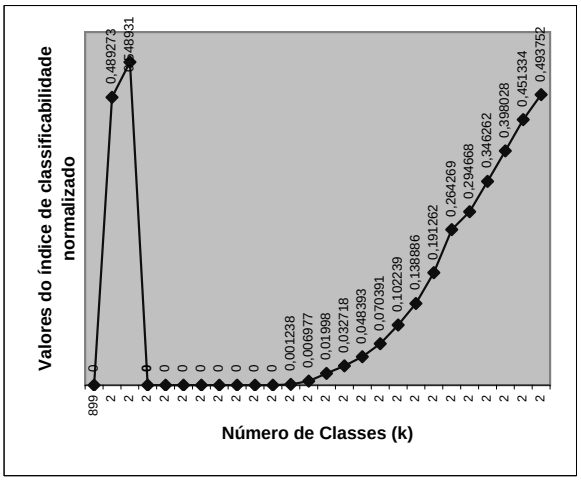

Figura 5.11: Gráfico $\tilde{IC}/k = 2$ obtido para o conjunto de dados representado na figura 5.9 com ajuste hierárquico.

do índice para cada uma das partições obtidas. A partição correspondente a $\alpha = 1$ não foi analisada porque é a partição obtida a partir do grafo completo e este não é representativo da estrutura dos dados, uma vez que o número de classes da partição obtida é aproximadamente igual ao número de elementos. Como se pode ver pela tabela, a partição seleccionada é a que corresponde a $\alpha = 7.52$, na qual as cinco classes são identificadas e 81 dos 100 pontos isolados são distribuídos por 64 classes. Como também se pode ver pela tabela, a partição seleccionada é a mais semelhante com a partição de referência. Nos casos do conjunto de formas não convexas representado na figura 4.7 e cujos resultados estão na quinta e sexta colunas da tabela, e do conjunto de dados com duas classes em forma de anel representado na figura 5.9 e cujos resultados estão na sétima e oitava colunas da tabela, opta-se por $k = 2$ uma vez que se obteve esta partição um maior número de vezes. Todas as partições em duas classes são iguais às partições de referência. A tabela 5.18 apresenta os resultados obtidos pelo ajuste hierárquico com k variável para 100 execuções de cada um dos conjuntos indicados.

Com os parâmetros apresentados na tabela 5.16, foram gerados 100 conjuntos de dados com 400 elementos cada, aos quais foram aplicados os diferentes métodos. Os resultados estão apresentados no Capítulo 8, no qual os resultados dos métodos MPCG são comparados com os obtidos por outros métodos de classificação.

Concluindo, para classes de formas não esféricas, o ajuste hierárquico apresenta um desempenho superior relativamente ao ajuste de densidade. No entanto, aquele ajuste

Tabela 5.17: Valores do índice IAP para as partições obtidas pelo ajuste hierárquico para k desconhecido para o conjunto de dados representado na figura 4.5.

α	k	IAP	IRC
7.52	69	0.823	0.971
14.03	46	0.733	0.950
20.55	20	0.631	0.907
27.06	7	0.622	0.525
33.58	2	0.611	0.295

Tabela 5.18: Média e desvio padrão dos valores de IRC para 100 execuções dos diferentes conjuntos de dados para ajuste hierárquico para k variável.

Conjunto de Dados	Média	Desvio Padrão
Figura 4.3	1	0
Figura 4.5	0.920	0.016
Figura 4.7	1	0

é menos eficaz na detecção de classes em conjuntos de dados com pontos isolados.

5.9 Conclusão

Neste capítulo foram apresentados alguns índices de avaliação de classes e partições, desenvolvidos no contexto da teoria de grafos e que, essencialmente, servem de suporte ao método proposto. Os índices de avaliação do isolamento absoluto e da densidade de classes permitem avaliar as classes obtidas pelo método; enquanto o índice de avaliação de partições, permite decidir qual a partição a seleccionar de entre aquelas que apresentem o mesmo valor do índice de classificabilidade. O índice de isolamento relativo serviu de suporte ao desenvolvimento de um ajuste hierárquico da coloração obtida pelo algoritmo sequencial, baseada em grafos ponderados pela medida de dissemelhança definida a partir do índice de isolamento. Esta abordagem mostrou-se eficiente na detecção de classes de forma alongada e de formas não convexas, superando o método das *k-médias*, o método hierárquico ascendente com o índice da média e a abordagem de ajuste de densidade adoptada inicialmente. Este método tem duas opções de funcionamento: k fixo ou k desconhecido. No caso do k fixo, o número de classes do conjunto de dados k, é atribuído *a priori*, enquanto que no caso de k desconhecido, a escolha da partição é feita, ou directamente pelos resultados

Tabela 5.19: Resultados da aplicação do ajuste hierárquico com α variável e k desconhecido.

Conjunto da Figura 4.3		Conjunto da Figura 4.5		Conjunto da Figura 4.7		Conjunto da Figura 5.9	
α	k	α	k	α	k	α	k
1	1368	1	878	1	490	1	899
3.08	4	7.52	69	6.80	25	3.79	100
5.16	4	14.03	46	12.59	3	6.57	3
7.25	4	20.55	20	18.38	2	9.36	2
9.33	4	27.06	7	24.18	2	12.14	2
11.41	4	33.58	2	29.97	2	14.93	2
13.49	3	40.10	1	35.76	2	17.72	2
15.58	3	46.61	1	41.56	2	20.50	2
17.66	3	53.13	1	47.35	2	23.29	2
19.74	3	59.65	1	53.14	2	26.07	2
21.82	1	66.16	1	58.94	1	28.86	2
23.91	1	72.68	1	64.73	1	31.64	1
25.99	1	79.19	1	70.53	1	34.43	1
28.07	1	85.71	1	76.32	1	37.22	1
30.15	1	92.23	1	82.11	1	40.00	1
32.24	1	98.74	1	87.91	1	42.79	1
34.32	1	105.26	1	93.70	1	45.57	1
36.40	1	111.77	1	99.49	1	48.36	1
38.48	1	118.29	1	105.29	1	51.15	1
40.57	1	124.81	1	111.08	1	53.93	1
42.65	1	131.32	1	116.88	1	56.72	1
44.73	1	137.84	1	122.67	1	59.50	1
46.81	1	144.36	1	128.46	1	62.29	1
48.89	1	150.87	1	134.26	1	65.07	1
50.98	1	157.39	1	140.05	1	67.86	1
53.06	1	163.91	1	145.85	1	70.65	1

do método (selecciona-se a partição que é identificada um maior número de vezes) ou usando o índice IAP. O ajuste hierárquico mostrou um melhor desempenho na identificação de classes de forma arbitrária, não sendo tão eficaz como o ajuste de densidade, na detecção de classes em conjuntos com pontos isolados.

Capítulo 6

Método com Atributos Diferenciados

Neste capítulo os atributos (ou variáveis) que caracterizam os dados vão ser considerados de modo diferenciado, quer quanto ao tipo, quer quanto ao modo como cada atributo participa no processo de classificação. A diferenciação dos atributos tem como suporte teórico os multigrafos, nos quais os arcos múltiplos entre vértices correspondem aos diferentes blocos de atributos. Vão ser apresentadas duas abordagens distintas com o objectivo de obter uma classificação não hierárquica no conjunto de dados. Numa primeira abordagem, são efectuadas projecções dos multigrafos em grafos simples, aos quais é aplicado o método de partição baseado na coloração de grafos (MPCG) ajustado, obtendo-se uma partição. Neste contexto, são apresentadas extensões dos índices definidos em capítulos anteriores aos atributos diferenciados. Uma segunda abordagem consiste em obter uma partição por cada bloco de atributos que, em conjunto, são usadas para formar uma partição consenso, representativa do conjunto de dados. Os métodos com atributos diferenciados vão ser aplicados a alguns conjuntos de dados reais, para os quais se põe em evidência alguma vantagem destas novas abordagens.

6.1 Atributos Diferenciados

Os atributos (ou variáveis) podem ser de tipo numérico ou categórico, sendo aplicadas funções de dissemelhança diferentes, adequadas a cada tipo de atributos.

6.1.1 Blocos de Atributos

Tal como na abordagem feita em [86], os atributos são agrupados em t_b blocos de acordo com o seu tipo, podendo ser numérico ou categórico. Designemos cada bloco com b_i e n_{b_i} o número de atributos do bloco b_i, para $i = 1, \ldots, t_b$. Todos os atributos pertencentes ao mesmo bloco são afectados com a mesma função de dissemelhança. No limite podemos ter $t_b = t$ blocos em que cada bloco é constituído por um único atributo.

6.1.2 Funções de Dissemelhança Específicas

As funções de dissemelhança específicas são definidas para cada bloco de atributos. A cada bloco de atributos pode ser associada uma função de dissemelhança diferente, seja ϕ_ℓ a função associada ao bloco b_ℓ para $\ell \in \{1, \ldots, t_b\}$.

Se o bloco de atributos b_ℓ for numérico então usa-se uma distância de *Minkowski*,

$$\phi_\ell^n(x_i^\ell, x_{i'}^\ell) = \left(\sum_{j=1}^{n_{b_j}} |x_{ij}^\ell - x_{i'j}^\ell|^q \right)^{\frac{1}{q}}$$

Em geral usa-se a distância euclideana obtida para $q = 2$. Se o bloco b_ℓ for categórico então

$$\phi_\ell^c(x_i^\ell, x_{i'}^\ell) = \sum_{j=1}^{n_{b_\ell}} \delta\left(x_{ij}^\ell, x_{i'j}^\ell\right) \tag{6.1}$$

onde

$$\delta\left(x_i^\ell, x_i'^\ell\right) = \begin{cases} 1 & \text{se } x_i^\ell \neq x_{i'}^\ell \\ 0 & \text{se } x_i^\ell = x_{i'}^\ell \end{cases} \tag{6.2}$$

6.1.3 Função de Dissemelhança Global

A função de dissemelhança global ϕ resulta da soma de t_b funções específicas, e mede a dissemelhança entre dois elementos x e y do conjunto Ω:

$$\phi(x,y) = \sum_{\ell=1}^{t_d} \beta^{\ell}\, \phi_{\ell}(x^{\ell}, y^{\ell})$$

sendo x^{ℓ} e y^{ℓ} os valores dos atributos correspondentes ao bloco b_{ℓ}, dos elementos x e y respectivamente.

A cada bloco de atributos é associado um peso de equilíbrio, permitindo minimizar o efeito que os diferentes domínios das variáveis podem ter no processo de classificação. Quando todos os atributos são numéricos, basta fazer uma estandardização das variáveis para que a compensação seja feita. No entanto, quando se combinam atributos numéricos com categóricos uma simples estandardização não é suficiente. Designe-se por β^{ℓ} o peso de equilíbrio associado ao bloco b_{ℓ}, em que $\ell \in \{1, \ldots, t_b\}$. Atribui-se ao peso de equilíbrio a associar aos atributos categóricos 30% da média aritmética dos desvios padrão dos atributos numéricos [111]. Verifica-se $\beta_{\ell} = 1$ se o bloco b^{ℓ} for numérico.

6.2 Multigrafo Representativo dos Dados

No multigrafo de suporte aos atributos diferenciados, a cada vértice corresponde um elemento do conjunto de dados. A unir dois vértices existe um arco múltiplo composto por, no máximo, t_b arcos simples, ou seja, no máximo, o número de arcos entre dois vértices é igual ao número de blocos de atributos.

No multigrafo existe um arco entre os vértices u e v, ao nível i se se verificar:

$$\phi_i(u,v) \geq \alpha_i \tag{6.3}$$

Neste caso, existe um parâmetro de controlo para cada bloco de atributos. Sejam $\alpha_1, \ldots, \alpha_{t_b}$ os t_b parâmetros de controlo.

Definição 6.1 *Seja $e_{ij}^m = \{v_i, v_j\}$ o arco múltiplo que une os vértices v_i e v_j. O* grau do arco múltiplo e_{ij}^m *designa-se por $g(e_{ij}^m)$ e é dado pelo número de arcos simples que compõem o arco múltiplo e_{ij}^m.*

Definição 6.2 *Seja V o conjunto dos vértices do multigrafo definido nos dados, no qual o conjunto de atributos está organizado em t_b blocos. Defina-se a função*

$$\eta : V \times V \mapsto \{1, \ldots, t_b\}$$

tal que

$$\eta(v_i, v_j) = g(e_{ij}^m) \tag{6.4}$$

onde $e_{ij}^m = \{v_i, v_j\}$, como sendo o número de arcos que unem os vértices v_i e v_j.

Designe-se por G o grafo projectado do multigrafo representativo dos dados com atributos diferenciados. O número de vértices mantém-se. No grafo projectado existe um arco entre dois vértices, se existirem pelo menos φ arcos a ligá-los no multigrafo, ou seja, estão presentes no grafo projectado todos os arcos múltiplos tais que $g(e^m) \geq \varphi$. O valor φ é uma constante fixada e designa-se por *força da ligação.*

Definição 6.3 *Seja φ um número tal que $1 \leq \varphi \leq t_b$, uma φ-projecção é um grafo simples com n vértices, no qual existe um arco a unir dois vértices, se existir em pelo menos φ arcos a ligar os vértices no multigrafo origem.*

Assim, no grafo projectado G, os vértices v_i e v_j são adjacentes se e só se

$$\eta(v_i, v_j) \geq \varphi$$

para φ fixado.

6.3 Extensão dos Índices aos Atributos Diferenciados

Usando o multigrafo representativo dos dados definido na secção anterior, nesta secção são estendidos o índice de classificabilidade (Secção 4.3, equação (4.17)), o índice de isolamento absoluto (Secção 5.3.1, equação (5.18)), o índice de isolamento relativo (Secção 5.4, equação (5.22)) e as duas componentes do índice de avaliação de partições, intraclasses (Secção 5.6, equação (5.34)) e entre classes (Secção 5.6, equação (5.36)), aos atributos diferenciados. Também para o ajuste hierárquico se vai fazer uma adaptação aos atributos diferenciados, optando-se por uma abordagem pela qual a escolha das classes a reunir em cada etapa, não é feita tomando apenas em conta a adjacência entre vértices, mas também considerando o número de arcos entre eles.

6.3.1 Índice de Classificabilidade

Recorde-se a fórmula (4.17) do índice de classificabilidade

$$\begin{aligned} IC &:= \frac{1}{2}\sum_{i=1}^{k}\sum_{v\in C_i}[d_G^{max}(v) - d_G(v)] \\ &= \frac{1}{2}\left[\sum_{i=1}^{k}\sum_{v\in C_i} d_G^{max}(v) - \sum_{i=1}^{k}\sum_{v\in C_i} d_G(v)\right] \end{aligned} \tag{6.5}$$

O grau máximo de $v \in C_i$ é $t_b(n - n_i)$, sendo t_b o número de blocos de atributos ou o número máximo de arcos entre dois vértices do multigrafo repesentativo dos dados, e n_i o número de elementos da classe C_i.

No contexto dos atributos diferenciados (AD), seja $V_\varphi^{AD}(v)$ a vizinhança de $v \in C_i$, ou seja,

$$V_\varphi^{AD}(v) = \{u \in V \setminus C_i : \eta(u, v) \geq \varphi\} \tag{6.6}$$

sendo φ a força da ligação entre os vértices do multigrafo. Designe-se por $d_{AD}(v)$ o grau do vértice v no contexto dos AD. O grau do vértice $v \in C_i$ é dado por:

$$d_{AD}(v) = \sum_{u\in V_\varphi^{AD}(v)} \eta(v, u) \tag{6.7}$$

A fórmula do índice de classificabilidade no contexto dos atributos mistos diferenciados é dado por:

$$\tilde{IC} := 1 - \frac{\displaystyle\sum_{i=1}^{k}\sum_{v\in C_i} d_{AD}(v)}{E_{max}^{AD}} \tag{6.8}$$

sendo E_{max}^{AD} o número máximo de arcos no multigrafo *k-partite* completo, dado por:

$$E_{max}^{AD} = \frac{t_b}{2}\left(n^2 - \sum_{i=1}^{k} n_i^2\right) \tag{6.9}$$

6.3.2 Índice de Isolamento Absoluto

Foi provado na secção 5.3.1 que o índice de isolamento absoluto normalizado é definido por:

$$\tilde{\Delta}_i^a = 1 - \frac{1}{n_i(n-n_i)} \sum_{v \in C_i} d_G(v)$$

no contexto dos AD o índice é difinido por:

$$\tilde{\Delta}_i^{AD} = 1 - \frac{1}{t_b n_i(n-n_i)} \sum_{v \in C_i} d_{AD}(v) \tag{6.10}$$

6.3.3 Índice de Isolamento Relativo

O índice de isolamento da classe C_i relativamente à classe C_j, no contexto do multigrafo definido nos dados, é dada por:

$$\tilde{\Delta}_i^{AD}(C_j) = 1 - \frac{1}{t_b n_i n_j} \sum_{v \in C_i} d_{AD}^{C_j}(v) \tag{6.11}$$

sendo

$$d_{AD}^{C_j}(v) = \sum_{u \in C_j \cap V_\varphi^{AD}(v)} \eta(v,u) \tag{6.12}$$

6.3.4 Índice de Avaliação de Partições

A parcela do índice de avaliação de partições referente aos vizinhos comuns de elementos pertencentes à mesma classe. Para cada uma das classes faz-se uma reordenação da matriz de adjacências, de modo a que os n_i elementos da classe C_i estejam nas n_i primeiras linhas da matriz. Assim, o valor da parcela IAP_{AD}^d é dado por:

$$IAP_{AD}^d = \frac{2}{k} \sum_{i=1}^{k} \left\{ \frac{\sum_{j=1}^{n_i-1} \sum_{\ell=j+1}^{n_i} a'_j \cdot a'_\ell}{t_b^2 n_i(n_i-1)(n-n_i)} \right\} \tag{6.13}$$

sendo a'_j e a'_ℓ as linhas j e ℓ da *matriz de adjacências múltiplas* $A'[a'_{ij}]$ do multigrafo. Esta é uma matriz $n \times n$ simétrica, tal que:

$$a'_{ij} = \begin{cases} \eta(v_i, v_j) & \text{se} \quad \eta(v_i, v_j) \geq \varphi \\ \\ 0 & \text{se} \quad \eta(v_i, v_j) < \varphi \end{cases} \tag{6.14}$$

Para cada par de classes C_i e C_j, faz-se uma reordenação da matriz de adjacências, de modo a que os n_i elementos da classe C_i estejam nas n_i primeiras linhas da matriz, estando os n_j elementos da classe C_j nas n_j linhas seguintes. Assim, a parcela do índice de avaliação de partições correspondente a vizinhos comuns entre elementos pertencentes a classes diferentes, representa-se por IAP^e_{AD} e é dada por:

$$IAP^e_{AD} = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \left\{ \frac{\sum_{p=1}^{n_i} \sum_{q=n_i+p}^{n_i+n_j} a'_p \cdot a'_q}{t_b^2 (n - n_i - n_j) n_i n_j} \right\} \tag{6.15}$$

Notar que, para todos os casos, se φ for igual ao número de atributos t, então as fórmulas (6.8), (6.10), (6.11), (6.13) e (6.15) são análogas às formulas apresentadas nos capítulos 4 e 5 para atributos não diferenciados.

6.3.5 Extensão do Ajuste Hierárquico

Recorde-se que no ajuste hierárquico, em cada etapa se reunem as classes com menor isolamento relativo, obtido pelo número de pares de vértices adjacentes entre as duas classes (Secção 5.8). No caso dos atributos diferenciados, a função de dissemelhança define-se do seguinte modo.

Definição 6.4 *Sejam C_i e C_j duas classes de n_i e n_j elementos, respectivamente, caracterizados por atributos diferenciados em t_b blocos. Define-se $\delta^{AD}(C_i, C_j)$ como uma medida de dissemelhança entre as classes do seguinte modo:*

$$\delta^{AD}(C_i, C_j) = \begin{cases} 0 & \textit{se} \quad i = j \\ \\ \dfrac{\sum_{u \in C_i} \sum_{v \in C_j} \eta(u, v)}{t_b n_i n_j} & \textit{se} \quad i \neq j \end{cases} \tag{6.16}$$

Assim, não se considera só se os vértices são ou não adjacentes ($\eta(u,v) \geq \varphi$), mas conta-se o número de arcos que os une. Quanto maior for esse número mais afastadas são as classes.

6.4 Partições em Grafos φ-Projectados

Nesta abordagem, o grafo representativo dos dados a ser usado pelo método de partição ajustado, resulta de uma φ-projecção do multigrafo obtido com atributos diferenciados, no qual existem, no máximo, t_b arcos a unir dois vértices. Designemos por *partições φ-projectadas*, as partições obtidas pelo aplicação do método ao grafo φ-projectado.

O modo de execução do método não se altera a não ser quanto à variação dos parâmetros de controlo. A partir do momento em que o método fixa todos os parâmetros, o modo de funcionamento é igual ao caso com atributos não diferenciados.

No caso dos atributos diferenciados com funções de dissemelhança específicas, considera-se

$$\ell^i_{min} \leq \alpha_i \leq \ell^i_{max}, \quad i = 1, \ldots, t_b$$

para cada bloco de atributos, sendo $\ell^i_{min} \geq \phi^{min}_i$ e $\ell^i_{max} \leq \phi^{max}_i$, onde ϕ^{min}_i e ϕ^{max}_i é a dissemelhança mínima e máxima, respectivamente, ao nível i (correspondendo ao bloco b_i), entre todos os pares de vértices do grafo. Tal como no caso dos atributos não diferenciados, os intervalos $[\ell^i_{min}, \ell^i_{max}]$ são divididos em H_i partes iguais.

Para uma enumeração completa, os parâmetros α_i, $i = 1, \ldots, t_b$, tomam todos os H_i valores possíveis, obtendo-se partições pelo algoritmo representado na tabela 6.1.

O procedimento descrito na tabela 6.2 determina o grafo φ-projectado G ao qual é aplicado o método de classificação MPCG, obtendo-se partições seguindo o mesmo procedimento que o seguido para o caso com atributos não diferenciados.

Seguindo esta abordagem, obtém-se $N = \Pi_{i=1}^{t_b} H_i$ partições. Das partições obtidas, seja N_k, $k \in \{1, \ldots, n\}$ o número de partições em k classes, sendo n o número de

Tabela 6.1: Algoritmo do método com atributos diferenciados para partições φ-projectadas.

Seja φ a força de ligação

Sejam $h_i = \dfrac{\ell^i_{max} - \ell^i_{min}}{H_i}$ $i = 1, \ldots, t_b$

Faça-se $\alpha_1 = \ell^1_{min}, \alpha_2 = \ell^2_{min}, \alpha_3 = \ell^3_{min}, \ldots, \alpha_{t_b} = \ell^{t_b}_{min}$

Enquanto $\alpha_1 \leq \ell^1_{max}$ faça-se

Enquanto $\alpha_2 \leq \ell^2_{max}$ faça-se

Enquanto $\alpha_3 \leq \ell^3_{max}$ faça-se

$\vdots$

Enquanto $\alpha_{t_b} \leq \phi^{max}_{t_b}$ faça-se

$G = encontra\_grafo\_projectado(\alpha_1, \alpha_2, \alpha_2, \ldots, \alpha_{t_b}, \varphi)$

$P = encontra\_particao\_MPCG(G_i)$

faça-se $\alpha_{t_b} = \alpha_{t_b} + h_{t_b}$

$\vdots$

faça-se $\alpha_3 = \alpha_3 + h_3$

faça-se $\alpha_2 = \alpha_2 + h_2$

faça-se $\alpha_1 = \alpha_1 + h_1$

Tabela 6.2: Algoritmo de obtenção do grafo φ-projectado.

Seja $MG(V, E)$ o multigrafo representativo do conjunto de dados

$G = encontra\_grafo\_projectado(\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_{t_b}, \varphi)$

$G(V) \leftarrow GM(V)$

$G(E) \leftarrow \emptyset$

para todos os pares $\{v_j, v_\ell\}$ faça

$n\_arcos_{j\ell} = \eta(v_j, v_\ell)$ em MG

se $n\_arcos_{j\ell} \geq \varphi$ então $G(E) = G(E) \cup \{\{v_j, v_\ell\}\}$

elementos a classificar. Verifica-se $N = \sum_{k=1}^{n} N_k$. Considerado k fixado, de entre as N_k partições em k classes, escolhe-se aquela que apresenta o maior valor do índice de avaliação de partições IAP (Secção 5.6). Os valores possíveis de k são aqueles para os quais foi gerada pelo menos uma partição em k classes.

A enumeração completa nos valores dos parâmetros de controlo, limita a aplicação desta abordagem a conjuntos de dados com não mais de quatro atributos.

6.5 Partições Consenso das Partições por Blocos de Atributos

Outra possível abordagem consiste em obter uma partição *consenso* das partições obtidas para cada bloco de atributos. Uma teoria formal para partições consenso é apresentada em [12] e uma revisão bibliográfica sobre o assunto é apresentada em [137].

6.5.1 Partições por Blocos de Atributos

Seja α_i o parâmetro de controlo associado ao bloco de atributos b_i, para $i = 1, \ldots, t_b$. A variação deste parâmetro é $\ell^i_{min} \leq \alpha_i \leq \ell^i_{max}$, sendo ℓ^i_{min} e ℓ^i_{max} a dissemelhança mínima e máxima, respectivamente, entre todos os pares de elementos, calculada apenas com os atributos pertencentes ao bloco b_i. O método de classificação MPCG aplica-se a cada bloco de atributos obtendo-se t_b partições, sejam $P = \{P_1, \ldots, P_{t_b}\}$.

Pretende-se que a partição representativa do conjunto de dados seja a partição consenso destas t_b partições.

6.5.2 Partição Consenso

Um dos processos possíveis para a obtenção de partições consenso consiste numa abordagem por optimização, pela qual se pretende minimizar uma função de controlo definida entre o conjunto das partições por blocos de atributos P e a partição consenso P_c [92].

Consideremos as t_b partições definidas por uma matriz tri-dimensional $C \equiv (c_{ij\ell})$, onde $c_{ij\ell} = 1$ (respectivamente, 0) se x_i, x_j pertencem (respectivamente, não pertencem) à mesma classe P_ℓ $(i, j = 1, \ldots, n;\ \ell = 1, \ldots, t_b)$. Seja $\Delta(P_r, P_\ell)$ uma medida da dissemelhança entre as partições P_r e P_ℓ. Uma definição comum para $\Delta(P_r, P_\ell)$ que foi usada em [92] e também vai ser adoptada, é o número de pares de objectos que pertencem à mesma classe numa das partições e a classes diferentes na outra:

$$\Delta(P_r, P_\ell) = \sum_{1 \leq i < j \leq n} (c_{ijr} - c_{ij\ell})^2 \tag{6.17}$$

A função de controlo consiste na soma dos valores de Δ entre cada partição $P_r \in P$ e a partição consenso. Outra possível definição para $\Delta(P_r, P_\ell)$, usada em [191] e mais tarde, por exemplo, em [101] e em [48], consiste no número mínimo de elementos que têm que ser transferidos entre classes em P_r, de tal forma que a partição resultante seja igual a P_ℓ. Em [133] é proposto o índice de *Rand* corrigido de *Hubert & Arabie* como medida entre partições a maximar para encontrar a partição consenso.

Pretende-se então minimizar a função de controlo (6.17), ou seja, pretende-se encontrar

a solução para o problema:

$$min \sum_{r=1}^{t_b} \Delta(P_r, P_c) \qquad \text{tal que} \qquad P_c \in \mathbb{P} \tag{6.18}$$

sendo $\mathbb{P}$ o conjunto de todas as partições possíveis do conjunto Ω de n elementos.

A solução do problema (6.18) consiste na obtenção de uma partição consenso *mediana* [92]. Esta partição foi inicialmente designada por partição central [191] e mais tarde designada por partição mediana por vários autores. Outros procedimentos da mediana aplicados a partições são apresentados em [14] e [13].

Outra possível abordagem consiste em encontrar a partição consenso *medoid* [92] pela qual a partição consenso é procurada de entre as partições observadas, ou seja

$$min \sum_{r=1}^{t_b} \Delta(P_r, P_c) \qquad \text{tal que} \qquad P_c \in P \tag{6.19}$$

Como o problema (6.18) é NP-Completo vai-se adoptar uma heurística baseada no número de vezes que um par de elementos é classificado junto nas t_b partições por blocos de atributos. Defina-se uma função de semelhança entre todos os pares de elementos de conjunto de dados, seja

$$s_{ij} = \sum_{r=1}^{t_b} \frac{c_{ijr}}{t_b}, \quad i, j = 1, \ldots, n \tag{6.20}$$

a proporção do número de vezes em que o par $\{x_i, x_j\}$ foi colocado na mesma classe nas partições $P_1, \ldots, P_{t_b}$. Tal como na abordagem apresentada em [213] e em [75], a função de semelhança (6.20) vai ser usada para se fazer uma reclassificação dos elementos do conjunto. Considere-se $d_{ij} = 1 - s_{ij}$. Esta função de dissemelhança associada a todos os pares de elementos do conjunto de dados, é usada no método de classificação MPCG para se obter a partição consenso.

Outras heurísticas podem ser usadas, por exemplo em [48], a heurística usada consiste em assumir que se um conjunto de elementos são sempre classificados juntos em todas

as partições iniciais, então também deverão ficar juntos na partição consenso. Em [14] é usada a heurística baseada no facto de que dois elementos estão na mesma classe na partição mediana, se e só se foram classificados na mesma classe em pelo menos $I/2 + 1$ ou $I/2 + 0.5$ das I partições iniciais.

O método de coloração sequencial foi aplicado com o ajuste de densidade e com ajuste hierárquico. Neste contexto, o parâmetro de controlo α poderá tomar qualquer valor entre zero e um.

Aplicando um algoritmo hierárquico ascendente com o índice da média sobre um conjunto de dados munido da função de semelhança definida em (6.20), e cortando a árvore ao nível 0.5, obtém-se uma boa solução para o problema (6.18) [92]. Assim, ao conjunto de dados munido da dissemelhança d, foi também aplicado o algoritmo de classificação ascendente hierárquico com o índice da média.

6.6 Aplicações

Segue-se a aplicação do método com atributos diferenciados a três conjuntos de dados reais.

6.6.1 Conjunto de Dados *Zoo*

O conjunto de dados *Zoo* foi obtido do *UCI Machine Learning Repository* [1] e foi originalmente criado por *Richard Forsyth* em 1990. É constituído por 101 animais caracterizados por 16 atributos, sendo 15 binários e 1 numérico. Os atributos caracterizam os animais com o objectivo de os classificar em um dos sete tipos: mamíferos (41 animais), pássaros (20 animais), répteis (5 animais), peixes (13 animais), anfíbios (4 animais), insectos (8 animais) e invertebrados (10 animais).

No caso do método com atributos diferenciados, em primeiro lugar, consideraram-se 16 blocos com um atributo cada, sendo 15 deles categóricos e um numérico; em segundo lugar, consideraram-se dois blocos tendo um deles os 15 atributos categóricos e o outro

[1] http://www.ics.uci.edu/∼mlearn/MLSummary.html.

Tabela 6.3: Resultados obtidos para o conjunto *Zoo* pelos métodos MPCG com atributos não diferenciados.

Método	α	$\tilde{IC}$	k	IRC	τ_a	IAP
MPCG / Ajuste de Densidade	1.614	0.002	19	0.729	0.992	0.560
MPCG / Ajuste Hierárquico (k = 7)	2.074	0.0005	7	0.808	0.945	0.622
MPCG / Ajuste Hierárquico (k desconhecido)	1.614	0	15	0.761	0.989	0.612

o atributo numérico. Na primeira abordagem não foi aplicado o método das partições φ-projectadas, devido ao elevado número de blocos.

6.6.1.1 Método com Atributos não Diferenciados

Para a aplicação do método com atributos não diferenciados, foi considerado um único bloco com os 16 atributos e como função global, a função dissemelhança entre dois elementos definida na Secção 3.9 pela Equação 3.20, sendo dada pela soma do quadrado da distância euclideana entre os atributos numéricos, e o número de categorias não coincidentes para cada um dos atributos categóricos.

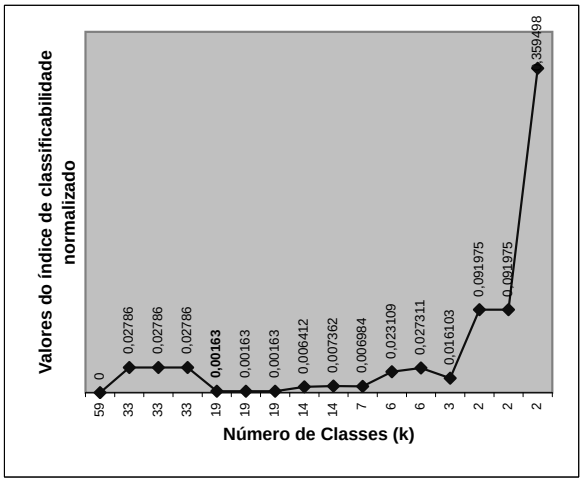

Figura 6.1: Gráfico $\tilde{IC}/k$ obtido pelo ajuste de densidade para o conjunto de dados *Zoo*.

O gráfico representado na figura 6.1 representa o resultado da aplicação do método MPCG com ajuste de densidade, obtendo-se uma partição em 19 classes com os valores associados $\alpha = 1.613575$ e $\tilde{IC} = 0.00163$. Como se pode ver na tabela 6.3, esta partição apresenta um elevado valor para o índice de validação $\tau_a = 0.992$, e um valor razoável para o índice de *Rand* corrigido, $IRC = 0.729$. No caso do ajuste hierárquico

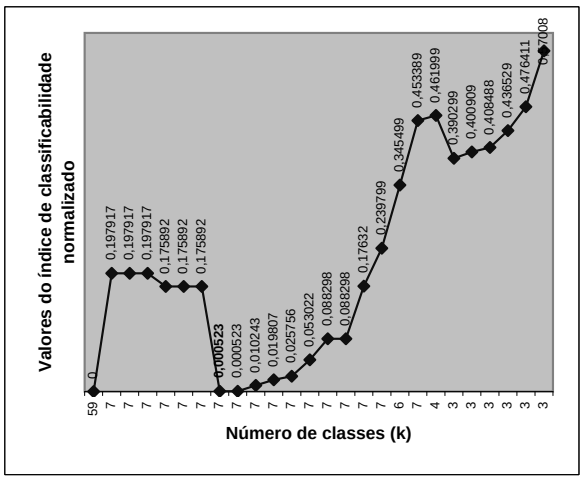

Figura 6.2: Gráfico $\tilde{IC}/k$ obtido pelo ajuste hierárquico para $k = 7$ para o conjunto de dados *Zoo*.

para $k = 7$, o resultado está representado na figura 6.2. A partição identificada pelo método apresenta os valores $\alpha = 2.073757$ e $\tilde{IC} = 0.000523$. Este valor do índice de classificabilidade é inferior ao obtido para o ajuste de densidade, identificando assim uma partição melhor, o que é constatado pelos valores do índice de $IRC = 0.808$. No entanto, de acordo com o critério do índice de validação τ_a esta última partição tem uma qualidade inferior à anterior, que apresenta um valor $\tau_a = 0.945$. No caso do ajuste hierárquico com k desconhecido, foi identificada uma partição em 15 classes para o valor $\alpha = 1.613575$, com uma partição mais semelhante à de referência do que a obtida pelo ajuste de densidade, mas menos semelhante que a obtida pelo ajuste hierárquico com $k = 7$.

6.6.1.2 Partições Consenso

Considerem-se em primeiro lugar $t_b = 16$ blocos com um atributo cada, sendo 15 deles categóricos e um numérico. Obtiveram-se as 16 partições, uma para cada bloco de atributos, cujos valores associados estão representados na tabela 6.4. No caso dos atributos binários consideram-se as partições definidas pelas categorias. No caso do atributo numérico (atributo 13) a partição obtida tem os valores associados: $\alpha = 0.96$, $\tilde{IC} = 0$ e $k = 6$.

Após a obtenção das dissemelhanças entre pares de elementos $d_{ij} = 1 - s_{ij}$, $i, j = 1, \ldots, n$ (Secção 6.5.2), aplicam-se os métodos de classificação MPCG fazendo variar

Tabela 6.4: Partições obtidas para cada um dos 16 blocos de atributos do conjunto de dados *Zoo*.

Bloco de Atributos	IRC	τ_a	IAP
1	0.383	0.714	0.334
2	0.455	0.472	0.330
3	0.562	0.694	0.333
4	**0.585**	**0.707**	**0.335**
5	0.406	0.492	0.332
6	0.352	0.656	0.332
7	0.184	0.590	0.333
8	0.504	0.638	0.334
9	0.392	0.731	0.330
10	0.382	0.720	0.332
11	0.277	0.798	0.347
12	0.256	0.740	0.330
13	0.298	0.921	0.431
14	0.201	0.667	0.334
15	0.073	0.676	0.347
16	0.129	0.629	0.332

α entre d_{min} e d_{max}, ou seja, entre a dissemelhança mínima e máxima de pares de elementos, respectivamente. De acordo com a definição da dissemelhança, para cada valor de α, consideram-se semelhantes os pares de elementos que ficaram na mesma classe em pelo menos $100 \times \alpha\%$ das partições por bloco de atributos. O gráfico representado na figura 6.3 representa o resultado da aplicação do método MPCG com ajuste de densidade, obtendo-se uma partição em 11 classes com os valores associados $\alpha = 0.26$ e $\tilde{IC} = 0.00223$. Como se pode ver na tabela 6.5, esta partição apresenta valores elevados para o índice de validação τ_a de *Guénoche* [98] (Secção 5.2), $\tau_a = 0.827$ e para o índice de *Rand* corrigido de *Hubert & Arabie* [113] (Secção 2.5.5), $IRC = 0.887$. No caso do ajuste hierárquico para $k = 7$, o resultado é representado na figura 6.4. A partição identificada pelo método apresenta os valores $\alpha = 0.245$ e $\tilde{IC} = 0.0005$.

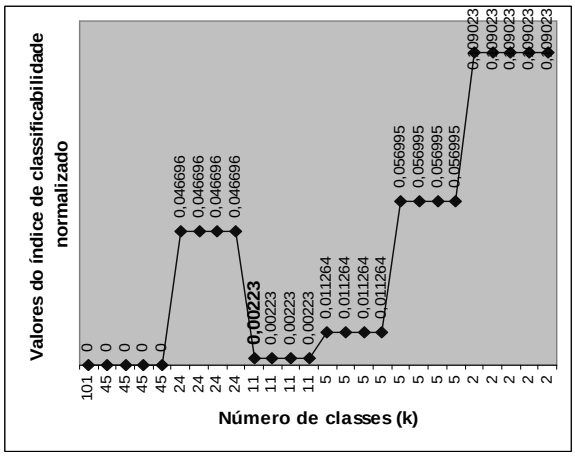

Figura 6.3: Gráfico $\tilde{IC}/k$ obtido pelo ajuste de densidade para o conjunto de dados *Zoo*, para a obtenção da partição consenso.

Como se pode ver pela tabela 6.5, os valores dos índices IAP e τ_a identificam como melhor partição consenso aquela que é obtida pelo ajuste hierárquico, tendo esta partição consenso um valor de IRC com a partição de referência, ligeiramente inferior ao que é obtido para a partição obtida pelo ajuste de densidade. Verifica-se que, para este exemplo, o método com atributos diferenciados encontrou uma partição mais semelhante ($IRC = 0.887$) com a de referência, que o método com atributos não diferenciados ($IRC = 0.808$).

A partição consenso *medoid* é dada pela partição associada ao quarto atributo (ver tabela 6.4), que como se pode ver pela tabela, também é a partição, de entre as 16, mais semelhante com a partição de referência.

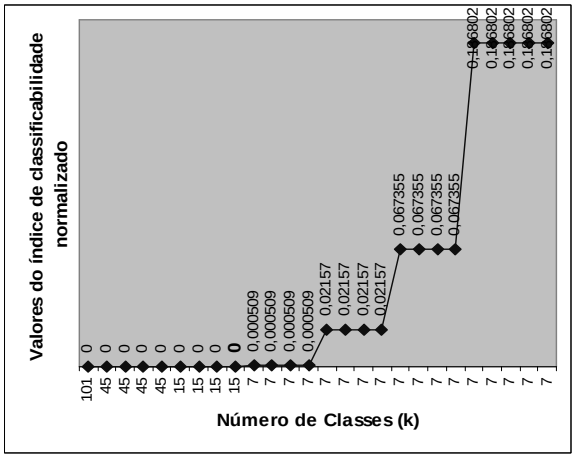

Figura 6.4: Gráfico $\tilde{IC}/k$ obtido pelo ajuste hierárquico com $k = 7$ para o conjunto de dados *Zoo*, para a obtenção da partição consenso.

Tabela 6.5: Partições consenso mediana obtidas pelos métodos MPCG, para o conjunto de dados *Zoo* para $t_b = 16$.

Partição Consenso Mediana	α	$\tilde{IC}$	k	IRC	τ_a	IAP
Ajuste de Densidade	0.260	0.002	11	0.887	0.827	0.482
Ajuste Hierárquico / k=7	0.245	0.0005	15	0.848	0.941	0.663
Ajuste Hierárquico / k desconhecido	0.245	0	15	0.848	0.941	0.663

Considerem-se de seguida os atributos agrupados em dois blocos, tendo um deles os 15 atributos categóricos, e o outro bloco, o atributo numérico. Designemos por *bloco categórico* o primeiro e *bloco numérico* o segundo. Obtiveram-se as 2 partições, uma para cada bloco de atributos, cujos valores associados estão representados na tabela 6.6.

Após a obtenção das dissemelhanças entre pares de elementos $d_{ij} = 1 - s_{ij}$, $i, j = 1, \ldots, n$ (Secção 6.5.2), aplicam-se os métodos de classificação MPCG para $\alpha = 0.5$,

Tabela 6.6: Partições obtidas para cada um dos dois blocos de atributos do conjunto de dados *Zoo*.

Bloco de Atributos	α	$\tilde{IC}$	k	IRC	τ_a	IAP
Categórico	**3**	**0.012**	**10**	**0.807**	**0.922**	**0.489**
Numérico	0.96	0	6	0.298	0.921	0.431

Tabela 6.7: Partições consenso mediana obtidas pelos métodos MPCG, para o conjunto de dados *Zoo*, para $t_b = 2$.

Partição Consenso Mediana	α	$\tilde{IC}$	k	IRC	τ_a	IAP
Ajuste de Densidade	0.5	0.0013	21	0.715	0.988	0.615
Ajuste Hierárquico / k=7	0.5	0	8	0.174	0.549	0.598

assim, consideram-se semelhantes os pares de elementos que ficaram na mesma classe em pelo menos uma das partições por bloco de atributos. A tabela 6.7 apresenta os resultados obtidos para os dois tipos de ajuste. Como se pode ver pela tabela, os valores dos índices IAP e τ_a identificam como melhor partição consenso aquela que é obtida pelo ajuste de densidade, tendo esta partição consenso um valor de IRC com a partição de referência, significativamente superior ao que é obtido para a partição obtida pelo ajuste hierárquico.

A partição consenso *medoid* é dada pela partição associada ao primeiro atributo (ver tabela 6.6), que como se pode ver pela tabela, também é a partição mais semelhante com a partição de referência.

Para este exemplo, a melhor partição ($IRC = 0.887$) foi a partição consenso mediana obtida pelo ajuste de densidade, quando aplicado ao conjunto com 16 blocos, um por atributo. No caso de se considerarem dois blocos, tendo um deles os atributos categóricos e o outro o atributo numérico, a partição consenso *medoid* ($IRC = 0.807$) é muito semelhante com a obtida pelo método com atributos não diferenciados e ajuste hierárquico ($IRC = 0.808$).

6.6.1.3 Partições φ-projectadas

Tal como no caso anterior, considerem-se os atributos agrupados em dois blocos, tendo um deles os 15 atributos categóricos e o outro o atributo numérico. Seja α_n o parâmetro de controlo associado ao bloco com o atributo numérico, considera-se $\phi_n^{min} \leq \alpha_n \leq \phi_n^{max}$ sendo ϕ_n^{min} e ϕ_n^{max} as dissemelhanças mínima e máxima, respectivamente, para o atributo numérico entre todos os pares de elementos. Seja α_c o parâmetro de controlo associado ao bloco com os atributos categóricos, considera-se $\alpha_c \in \{1, 2, \ldots, 15\}$ uma vez que se trata de atributos categórios, para este caso, a dissemelhança entre dois elementos varia entre 0 e 15.

Tabela 6.8: Partições φ-projectadas obtidas pelos métodos MPCG, para o conjunto de dados *Zoo*, para $\varphi = 2$ e $t_b = 2$.

Métodos	α_n	α_c	$\tilde{IC}$	k	IRC	τ_a	IAP
Ajuste de Densidade	1	1	0.005	6	0.514	0.921	0.947
Ajuste Hierárquico / k=7	1	5	0.264	7	0.443	0.892	0.616
Ajuste Hierárquico / k desconhecido	1	5	0	7	0.443	0.892	0.616

Tabela 6.9: Partições φ-projectadas obtidas pelos métodos MPCG, para o conjunto de dados *Zoo*, para $\varphi = 1$ e $t_b = 2$.

Métodos	α_n	α_c	$\tilde{IC}$	k	IRC	τ_a	IAP
Ajuste de Densidade	1	1	0.002	6	0.511	0.903	0.919
Ajuste Hierárquico / k=7	1	1	0.006	7	0.517	0.923	0.879
Ajuste Hierárquico / k desconhecido	1	1	0	7	0.517	0.923	0.879

As tabelas 6.8 e 6.9 apresentam os resultados obtidos, para $\varphi = 2$ e $\varphi = 1$, respectivamente. Recorde-se que $\varphi = 1, \ldots, t_b$ sendo t_b o número de blocos (Secção 6.4).

Como se pode constatar pela análise das tabelas, os resultantes foram muito semelhantes entre si. De acordo com o índice τ_a, a melhor partição é a de $k = 7$ classes obtida pelos ajustes hierárquicos para $\varphi = 1$, esta partição apresenta os valores $IAP = 0.879$ e $IRC = 0.517$. De acordo com o índice IAP a partição a selecionar seria a de $k = 6$ classes obtida pelo método com ajuste de densidade e $\varphi = 2$, esta partição apresenta os valores $\tau_a = 0.921$ e $IRC = 0.514$, valores estes muito semelhantes com os anteriores. Finalmente, a índice IRC indica que a partição identificada pelo índice τ_a é a mais semelhante com a partição de referência.

6.6.2 Conjunto de Dados *Iris*

O conjunto de dados *Iris* [214] é constituído por 150 flores, 50 de cada uma de três variedades de Iris: Setosa, Versicolor e Virginica; caracterizadas por quatro atributos numéricos que descrevem a largura e o comprimento das pétalas e das sépalas de cada tipo de flores. As flores são numeradas de 1 a 150, sendo as x_1 a x_{50} flores do tipo Iris Setosa, as x_{51} a x_{100} flores do tipo Iris Versicolor e finalmente as x_{101} a x_{150} flores do

Tabela 6.10: Resultados obtidos para o conjunto de dados *Iris*, pelos métodos MPCG com atributos não diferenciados.

Métodos	α	$\tilde{IC}$	k	IRC	τ_a	IAP
Ajuste de Densidade	1.218	0	2	0.568	0.929	1
Ajuste Hierárquico / k=3	0.798	0	3	0.564	0.930	0.827
Ajuste Hierárquico / k desconhecido	1.218	0	2	0.568	0.929	1

tipo Iris Virginica. Os atributos apresentam-se pela seguinte ordem: (1) comprimento da sépala, (2) largura da sépala, (3) comprimento da pétala e (4) largura da pétala, sendo todas as quantidades medidas em centímetros. A partição de referência para este conjunto de dados indica que uma das classes (Iris Setosa) é bem separada das outras duas, que se intersectam.

No caso do método com atributos diferenciados, vão ser consideradas duas situações: (1) formação de 4 blocos com um atributo cada e (2) formação de dois blocos com dois atributos cada; o primeiro bloco tem os atributos 1 e 2 referentes às características das sépalas, e o segundo bloco tem os atributos 3 e 4 referente às características das pétalas.

6.6.2.1 Método com Atributos não Diferenciados

Para a aplicação do método com atributos não diferenciados foi considerado um único bloco com os quatro atributos e como função global, a função distância euclideana.

A figura 6.5 representam o gráfico obtido da aplicação do método MPCG com ajuste de densidade, da qual resultou a identificação de uma partição em duas classes com os valores associados de $\alpha = 1.217631$, $\tilde{IC} = 0$ e $k = 2$. Esta partição consiste na classe Setosa isolada, estando as outras duas fundidas numa só. Esta partição apresenta os valores $\tau_a = 0.929$, $IRC = 0.568$ com a partição de referência e $IAP = 1$. De acordo com o índice de validação τ_a, esta partição é de melhor qualidade que a de referência que apresenta um valor de $\tau_a = 0.797$.

No caso do ajuste hierárquico com $k = 3$ obteve-se o gráfico representado na figura 6.6, pelo qual se identifica a partição em três classes para os valores associados de $\alpha = 0.798519$, $\tilde{IC} = 0$ e $k = 3$. No entanto, apresenta um valor $IRC = 0.564$ não correspondendo à separação das floras nas três classes de 50 flores cada. Esta partição

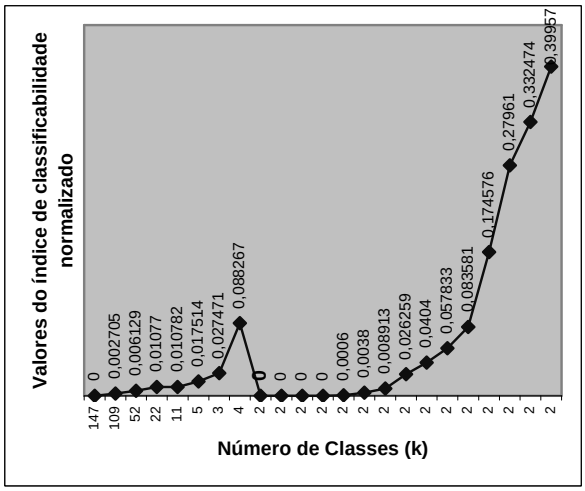

Figura 6.5: Gráfico $I\tilde{C}/k$ obtido pelo ajuste de densidade para o conjunto de dados *Iris*.

identifica separadamente as flores Iris Setosa num classe, considera duas flores do tipo Iris Virginica noutra classe, colocando as restantes numa única classe. Esta partição apresenta os valores $\tau_a = 0.930$ e $IAP = 0.827$.

No caso da aplicação do método MPCG com ajuste hierárquico e k desconhecido obteve-se o mesmo resultado que para o caso com ajuste de densidade. A tabela 6.10 apresenta os resultados obtidos para estes três métodos.

Vejamos se o método com atributos diferenciados consegue encontrar partições mais idênticas à de referência ou com valores dos índices de validação mais elevados.

6.6.2.2 Partições φ-projectadas: $t_b = 4$

Consideraram-se quatro blocos com um atributo numérico cada. Seja α_i o parâmetro de controlo associado ao atributo i, considera-se $\phi_i^{min} \leq \alpha_i \leq \phi_i^{max}$ para $i = 1, \ldots, 4$ e sendo ϕ_i^{min} e ϕ_i^{max} as dissemelhanças mínima e máxima, respectivamente, para o atributo i entre todos os pares de elementos, respectivamente.

Para a força de ligação $\varphi = 4$, obtiveram-se partições em $k = 7, 6, 5, 4, 3, 2$ classes, sendo em maior número (88% dos casos), as partições em duas classes. De entre as partições obtidas pelo método em $k = i$ classes, $(i = 2, \ldots, 7)$ escolheu-se aquela

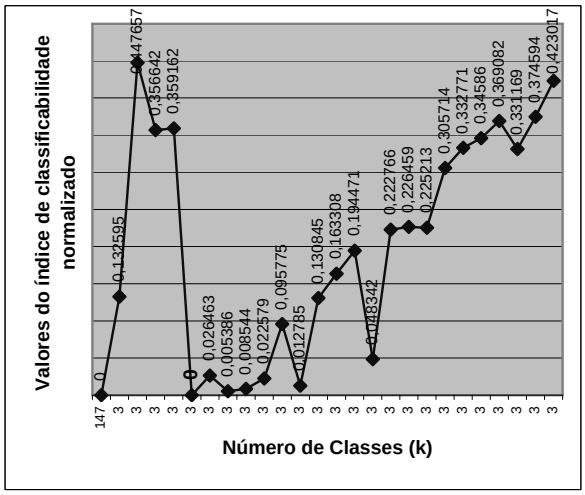

Figura 6.6: Gráfico $\tilde{IC}/k$ obtido pelo ajuste hierárquico para $k = 3$, para o conjunto de dados *Iris*.

Tabela 6.11: Resultados da aplicação do método das partições φ-projectadas ao conjunto de dados *Iris*, para $\varphi = 4$ e $t_b = 4$.

k	α_1	α_2	α_3	α_4	IRC	τ_a	IAP
7	0.38	0.33	0.59	0.24	0.490	0.843	0.390
6	0.38	0.514	0.472	0.24	0.478	0.816	0.326
5	0.45	0.284	0.472	0.624	0.859	0.858	0.513
4	0.38	0.284	1.534	0.528	0.794	0.871	0.471
3	1.01	0.33	0.944	0.24	0.504	0.866	0.443
2	0.45	0.284	2.242	0.528	0.540	0.935	0.400

partição com o maior valor do índice de avaliação de partições. Os parâmetros de controlo que originaram as partições seleccionadas e os respectivos valores do índice IRC com a partição de referência estão representados na tabela 6.11.

Como se pode concluir pela análise da tabela, com o método com atributos diferenciados, o índice IAP identificou uma partição mais semelhante com a de referência do que a que tinha sido identificada pelo método dos atributos não diferenciados. Essa partição em cinco classes é composta por uma classe com as 50 flores do tipo Setosa, outra classe é composta por uma flôr do tipo Virginica, a terceira classe é composta por duas flores também do tipo Virginica, a outra classe é composta por 48 flores do tipo Versicolor mais quatro flores do tipo Virginica, finalmente a quinta classe é composta por 43 flores do tipo Virginica e duas flores do tipo Versicolor. Esta partição

Tabela 6.12: Partições obtidas para cada um dos quatro blocos de atributos para o conjunto de dados *Iris*.

Atributo	α	$\tilde{IC}$	k	IRC	τ_a	IAP
1	0.87	0.323	5	0.324	0.802	0.292
2	1	0.323	3	0.156	0.582	0.204
3	2.596	0.234	3	0.508	0.917	0.247
4	**0.624**	**0.169**	**4**	**0.837**	**0.850**	**0.297**

apresenta um valor $\tau_a = 0.858$.

No caso da aplicação do método com força de ligação $\varphi = 3$ e $\varphi = 2$ obtiveram-se as partições em duas e três classes já identificadas pelo método com ajuste de densidade e hierárquico para atributos não diferenciados. Para o caso de $\varphi = 1$, para todas as combinações dos parâmetros de controlo, obtiveram-se partições com uma única classe.

6.6.2.3 Partições Consenso: $t_b = 4$

No caso do método com atributos diferenciados com partições consenso, consideraram-se quatro blocos com um atributo cada, e obtiveram-se as quatro partições, correspondentes aos quatro atributos, cujos valores associados estão representados na tabela 6.12.

Após a obtenção das dissemelhanças entre pares de elementos $d_{ij} = 1 - s_{ij}$, $i, j = 1, \ldots, n$ (Secção 6.5.2), aplicam-se os métodos de classificação MPCG para $\alpha = 0.5$, ou seja, considerando-se semelhantes os pares de elementos que ficaram na mesma classe em pelo menos 50% das partições por bloco de atributos. A tabela 6.13 apresenta os resultados obtidos para os dois tipos de ajuste. Como se pode ver pela tabela, os valores dos índices IAP e τ_a identificam como melhor partição consenso aquela que é obtida pelo ajuste hierárquico, tendo também esta partição consenso um valor de IRC com a partição de referência, superior ao que é obtido para a partição obtida pelo ajuste de densidade.

Finalmente a partição consenso *medoid* é dada pela partição associada ao quarto atributo (ver tabela 6.12), que como se pode ver pela tabela, também foi considerada

Tabela 6.13: Partições consenso mediana obtidas pelo método MPCG, para o conjunto de dados *Iris* para $t_b = 4$.

Partição Consenso	α	$\tilde{IC}$	k	IRC	τ_a	IAP
Ajuste de Densidade	0.5	0.054	4	0.525	0.827	0.351
Ajuste Hierárquico / k=3	0.5	0.001	3	0.560	0.931	0.403

Tabela 6.14: Resultados da aplicação do método das partições φ-projectadas ao conjunto de dados *Iris*, para $\varphi = 2$ e $t_b = 2$.

Método	$\tilde{IC}$	k	α_1	α_2	IRC	τ_a	IAP
Ajuste de Densidade	0.05	3	0.679	1.002	0.560	0.926	0.891
Ajuste de Densidade	0.0858	2	0.823	2.004	0.568	0.929	0.853
Ajuste Hierárquico / k=3	0.051	3	0.679	1.503	0.560	0.934	0.867

a melhor partição de entre as quatro dos atributos, de acordo com o índice IAP.

6.6.2.4 Partições φ-projectadas: $t_b = 2$

Considerem-se dois blocos com dois atributo numérico cada, estando os atributos 1 e 2 num bloco e os outros dois no outro bloco. Seja α_i o parâmetro de controlo associado ao atributo i, considera-se $\phi_i^{min} \leq \alpha_i \leq \phi_i^{max}$ para $i = 1, 2$ e sendo ϕ_i^{min} e ϕ_i^{max} as dissemelhanças mínima e máxima, respectivamente, para o bloco de atributos i entre todos os pares de elementos, respectivamente.

As tabelas 6.14 e 6.15 apresentam os resultados obtidos para $\varphi = 2$ e $\varphi = 1$ respectivamente.

Como se pode ver pelas tabelas, as partições φ-projectadas aplicadas aos atributos agrupados em dois blocos, não melhorou o desempenho dos métodos relativamente à

Tabela 6.15: Resultados da aplicação do método das partições φ-projectadas ao conjunto de dados *Iris*, para $\varphi = 1$ e $t_b = 2$.

Método	$\tilde{IC}$	k	α_1	α_2	IRC	τ_a	IAP
Ajuste de Densidade	0	2	0.245	0.501	0.568	0.929	1
Ajuste Hierárquico / k=3	0.010	3	0.389	1.253	0.561	0.923	0.875

Tabela 6.16: Partições obtidas para cada um dos dois blocos de atributos para o conjunto de dados *Iris*.

Bloco de Atributos	α	$\tilde{IC}$	k	IRC	τ_a	IAP
1	0.606	0.017	4	0.526	0.905	0.589
2	0.877	0	2	0.568	0.929	0.500

Tabela 6.17: Partições consenso mediana obtidas pelos métodos MPCG, para o conjunto de dados *Iris*, para $t_b = 2$.

Partição Consenso Mediana	α	$\tilde{IC}$	k	IRC	τ_a	IAP
Ajuste de Densidade	0.5	0.017	6	0.542	0.905	0.589
Ajuste Hierárquico / k=3	0.6	0.04	2	0.568	0.929	0.608

abordagem com atruibutos não diferenciados. A partição mais semelhante com a de referência é a obtida pela aplicação do ajuste de densidade para $\varphi = 1$ ($IRC = 0.568$).

6.6.2.5 Partições Consenso: $t_b = 2$

A tabela 6.16 apresenta as duas partições, correspondentes aos dois blocos de atributos, identificadas pelos método.

Após a obtenção das dissemelhanças entre pares de elementos $d_{ij} = 1 - s_{ij}$, $i, j = 1, \ldots, n$ (Secção 6.5.2), aplicam-se os métodos de classificação MPCG para $\alpha = 0.5$, ou seja, considerando-se semelhantes os pares de elementos que ficaram na mesma classe em pelo menos 50% das partições por bloco de atributos. A tabela 6.17 apresenta os resultados obtidos para os dois tipos de ajuste. Como se pode ver pela tabela, os valores dos índices IAP e τ_a identificam como melhor partição consenso aquela que é obtida pelo ajuste hierárquico, tendo também esta partição consenso um valor de IRC com a partição de referência, superior ao que é obtido para a partição obtida pelo ajuste de densidade.

No caso do ajuste hierárquico optou-se por $\alpha = 0.6$ porque com o valor $\alpha = 0.5$, a partição obtida consistia em três classes, estando 147 dos elementos numa única classe.

Finalmente, quanto à partição consenso *medoid* o critério de selecção da partição

obteve o mesmo valor para as duas partições por blocos de atributos.

Com o método para atributos diferenciados considerando dois blocos com dois atributos cada, não se obtiveram melhores resultados do que os obtidos pelos métodos com atributos não diferenciados.

A tabela 6.18 apresenta as partições candidatas à melhor partição de acordo com os índices e métodos utilizados. A partição P_1 foi a obtida pelo método com ajuste de densidade e atributos não diferenciados, pelo método das partições φ-projectadas para $\varphi = 1$ e $\varphi = 2$ e $t_b = 2$, e foi ainda obtida pelo método de conseno mediana com ajuste hierárquico para $t_b = 2$. A partição P_2 foi obtida pelo método com ajuste hierárquico e atributos não diferenciados. A partição P_3 é a candidata das partições φ-projectadas ($\varphi = 4$ e $t_b = 4$), sendo considerada a melhor das indicadas na tabela 6.11 pelo índice IAP. Para este caso, o índice de validação τ_a considerou como a melhor partição a composta por duas classes, no entanto, como nesta partição a classe Setosa não foi identificada correctamente, consideramos somente a partição de cinco classes. A partição P_4 é a partição *medoid* consenso correspondente ao quarto atributo (tabela 6.12), finalmente a partição P_5 foi considerada a melhor pelos dois índices, de entre as partições de consenso obtidas pelos métodos com ajuste de densidade e hierárquico (tabela 6.13). Como se pode ver pela análise da tabela, as melhores partições são as que consideram que o conjunto de dados é composto por duas classes, que é o caso da partição P_1 e P_2 (em que uma das classes é composta por dois elementos). A terceira melhor partição de acordo com o índice IAP é a partição P_3 que é aquela entre todas, é mais semelhante com a partição de referência. No caso do índice de validação τ_a, considerou-se como melhor a partição consenso, obtida para o ajuste hierárquico, a qual identifica correctamente a classe Setosa, considerando seis flores Virginica numa classe, ficando todas as restantes numa terceira classe. Para este índice, a quarta melhor partição é a correspondente à partição mais semelhante com a de referência.

6.6.3 Conjunto de Dados da UNESCO

O conjunto de dados que vai ser analisado nesta secção foi obtido da página da Internet da UNESCO [2] (*United Nations Educational, Scientific and Cultural Organization*), e consiste nos racios (*net enrolment ratio*) entre o número de jovens entre os 11 e os 15

[2] http://www.uis.unesco.org

Tabela 6.18: Partições obtidas pelo método de classificação para o conjunto de dados *Iris*.

Partições	k	τ_a	IAP	IRC
P_1	2	0.929	1	0.568
P_2	3	0.930	0.827	0.564
P_3	5	0.858	0.513	0.859
P_4	4	0.850	0.297	0.837
P_5	3	0.931	0.403	0.560

anos matriculados em instituições do ensino primário e secundário público e privado, e o número total de indivíduos da população nessa faixa etária, de alguns países. Os racios que excedem os 100% reflectem discrepâncias entre os dois conjuntos de valores.

Foram analisados 45 países e as tabelas 6.19 e 6.20 apresentam os resultados obtidos. A primeira engloba os países participantes no programa WEI (*World Education Indicators*), enquanto a segunda tabela apresenta os resultados obtidos para os países da OECD (*Organisation for Economic Co-operation and Development*). No presente estudo, as duas tabelas foram consideradas em conjunto. Os dados da Argentina, Brasil, Chile, Peru, Malásia e Filipinas referem-se ao ano 1998, os dados da Indonésia referem-se ao ano 2000, todos os restantes referem-se ao ano de 1999.

Note-se que para este exemplo não há partição de referência.

6.6.3.1 Método com Atributos não Diferenciados

Tal como foi feito para os exemplos anteriores, foi considerado um único bloco com os cinco atributos e como função global, a função distância euclideana.

A figura 6.7 representa o gráfico obtido da aplicação do método MPCG com ajuste de densidade, da qual resultou a identificação de uma partição em sete classes com os valores associados de $\alpha = 21.246$, $\tilde{IC} = 0$ e $k = 7$. Esta partição apresenta os valores $\tau_a = 0.939$ e $IAP = 0.676$. No caso do ajuste hierárquico obtiveram-se resultados semelhantes.

Tabela 6.19: Frequência de escolaridade primária e secundária de indivíduos entre os 11 e os 15 anos nos países participantes WEI (Fonte: UNESCO).

		11 anos	12 anos	13 anos	14 anos	15 anos
1	Argentina	106	105	95	91	80
2	Brasil	97	97	96	93	85
3	Chile	95	93	90	89	86
4	Egipto	93	86	73	65	67
5	Indonésia	79	94	70	59	45
6	Jordânia	87	85	87	79	79
7	Malásia	95	100	92	90	76
8	Paraguai	91	90	78	70	59
9	Peru	105	99	95	86	82
10	Filipinas	95	61	82	78	80
11	Federação Russa	90	95	97	89	74
12	Tailândia	99	100	93	86	77
13	Tunísia	102	94	91	84	71
14	Uruguai	109	129	99	81	72
15	Zimbabwe	93	93	121	39	51
	Médias WEI	**9**5.7	**9**4.7	**9**0.6	**7**8.5	**7**2.3

Tabela 6.20: Frequência de escolaridade primária e secundária de indivíduos entre os 11 e os 15 anos nos países da OECD (Fonte: UNESCO).

		11 anos	12 anos	13 anos	14 anos	15 anos
16	Austrália	99	99	98	99	97
17	Áustria	99	99	99	99	95
18	Bélgica	98	99	99	99	100
19	Belgica (Fl)	96	96	96	95	97
20	Canadá	98	99	98	97	98
21	República Checa	100	100	100	100	100
22	Dinamarca	100	100	100	98	97
23	Finlândia	99	99	99	99	100
24	França	99	99	99	99	98
25	Alemanha	99	99	100	100	99
26	Grécia	99	99	96	97	93
27	Hungria	101	101	99	99	96
28	Islândia	99	99	99	100	98
29	Irlanda	101	99	100	99	102
30	Itália	99	101	99	94	88
31	Japão	101	103	102	101	99
32	Coreia	99	94	100	98	97
33	Luxemburgo	98	90	99	95	92
34	México	96	90	82	73	55
35	Holanda	98	99	100	99	102
36	Nova Zelândia	100	99	98	98	96
37	Noruega	99	99	99	99	100
38	Polónia	97	97	97	97	96
39	Portugal	110	115	111	112	95
40	Espanha	106	105	105	104	96
41	Suécia	101	101	101	101	97
42	Suiça	100	100	99	99	97
43	Turquia	89	74	63	51	47
44	Grã-Bretanha	99	99	98	98	103
45	Estados Unidos	104	113	102	96	107
	Médias OECD	**9**9.4	**9**8.9	**9**7.9	**9**6.5	**9**4.6

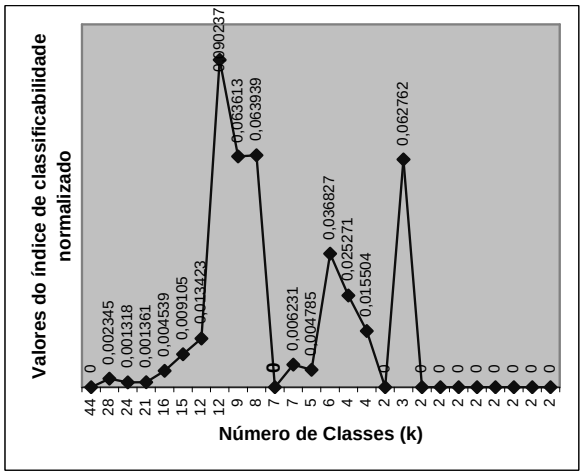

Figura 6.7: Gráfico $\tilde{IC}/k$ obtido pelo ajuste de densidade para o conjunto de dados UNESCO.

Tabela 6.21: Partições obtidas para cada um dos cinco blocos de atributos do conjunto de dados da UNESCO.

Blocos de Atributos	α	$\tilde{IC}$	k	τ_a	IAP
1	8.2	0.525	4	0.743	0.222
2	3.68	0.104	9	0.792	0.519
3	13.54	0.382	5	0.764	0.342
4	**12.52**	**0.312**	**6**	**0.931**	**0.405**
5	5.88	0.114	10	0.865	0.387

Esta partição considerou como pontos isolados os países: Indonésia (5), Filipinas (10), Uruguai (14), Zimbabwe (15) e Turquia (43), que como se pode ver pela tabela de dados, apresentam valores não usuais. Outra classe é constituída por três países: México (34), Egipto (4) e Paraguai (8) que apresentam valores semelhantes. Finalmente os restantes países estão colocados numa única classe.

6.6.3.2 Partições Consenso

No caso do método com atributos diferenciados com partições consenso, consideraram-se cinco blocos com um atributo cada ($t_b = 5$), e obtiveram-se as cinco partições cujos valores associados estão representados na tabela 6.21.

Tabela 6.22: Partição consenso mediana obtida pelo método MPCG para o conjunto de dados da UNESCO.

Partição Consenso Mediana	α	$\tilde{IC}$	k	τ_a	IAP
Ajuste de Densidade	0.5	0.005	9	0.909	0.732

Após a obtenção das dissemelhanças entre pares de elementos $d_{ij} = 1 - s_{ij}$, $i, j = 1, \ldots, n$ (Secção 6.5.2), aplicou-se o método de classificação MPCG para $\alpha = 0.5$. Como para este caso não há partição de referência, desconhecendo-se o valor de k, não se aplicou nenhum dos métodos hierárquicos. A tabela 6.22 apresenta os resultados obtidos para o ajuste de densidade.

A partição consenso obtida para o ajuste de densidade é semelhante à que foi obtida para atributos não diferenciados. Relativamente a países colocados isoladamente considerou: Indonésia (5), Uruguai (14), Zimbabwe (15), Jordânia (6), Turquia (43) e Portugal (39), que como se pode ver pela tabela de valores apresentam valores não usuais. São formadas duas classes com dois países cada: Egipto (4) e Paraguai (8) numa classe e Filipinas (10) e México (34) noutra classe. Finalmente os restantes países estão agrupados numa única classe.

A partição consenso *medoid* é dada pela partição associada ao quarto atributo (ver tabela 6.21), que apresenta um valor $\tau_a = 0.931$ e $IAP = 0.405$.

Pela análise dos resultados, e segundo o índice de validação de partições τ_a, a melhor partição corresponde à obtida pelo método com ajuste de densidade aplicado a atributos não diferenciados ($\tau_a = 0.939$). Segundo o mesmo índice, a segunda melhor partição corresponde à obtida para o quarto atributo ($\tau_a = 0.931$). De acordo com o índice IAP, a melhor partição consenso é a obtida para o método com atributos diferenciados usando partições consenso e ajuste de densidade ($IAP = 0.732$), sendo a segunda melhor, obtida para os atributos não diferenciados ($IAP = 0.676$). Analisando os resultados pela tabela de dados, verifica-se que a partição identificada pelo índice IAP é ligeiramente "melhor" do que a identificada pelo índice τ_a, uma vez que aquela considerou como pontos isolados, dois dos países (Jordânia (6) e Portugal (39)) com valores não usuais, que não foram isolados na partição com atributos não diferenciados.

6.7 Conclusão

Este capítulo apresentou uma nova abordagem do método, pelo qual as partições não são obtidas usando todos os atributos em simultâneo, mas por partições encontradas em sub-espaços de atributos. A diferenciação dos atributos permite que os atributos ou blocos de atributos tenham um contributo individual para a identificação de partições no conjunto de dados. O método das partições φ-projectadas mostrou-se eficiente na detecção da partição de referência para o conjunto de dados *Iris*, para $\varphi = 4$ e $t_b = 4$, apresentando um valor de $IRC = 0.859$ com a partição de referência. Devido a uma enumeração completa nos valores dos parâmetros de controlo, esta abordagem é computacionalmente muito pesada pelo que se restringe a sua aplicação a conjuntos de dados com, no máximo, quatro atributos. Outra abordagem consiste em identificar partições de consenso das partições obtidas para cada bloco de atributos. No caso do conjunto de dados *Zoo*, a partição mais semelhante com a de referência, foi a partição consenso mediana obtida com ajuste de densidade para $t_b = 16$ ($IRC = 0.887$). Para este conjunto de dados, os atributos foram ainda também agrupados em dois blocos ($t_b = 2$), tendo um deles os atributos categóricos e o outro o atributo numérico; no entanto, com esta abordagem não foi identificada uma partição melhor do que para o caso anterior. A abordagem das partições consenso também se mostrou eficiente na detecção da partição do conjunto *Iris*, apresentando um valor $IRC = 0.837$ com a partição de referência.

Capítulo 7

Estudo Estatístico dos Índices usando Teoria de Grafos Aleatórios

A avaliação do valor de uma medida de validação de uma estrutura em classes, exige uma comparação com o valor do índice obtido sob um modelo nulo de ausência de estrutura. A ideia base é testar se os elementos do conjunto de dados estão uniformemente estruturados. Esta análise é baseada na *hipótese nula*, H_0, expressa como uma afirmação de uma estrutura uniforme nos dados. A hipótese nula deverá ser testada comparando o valor da estatística de teste, obtida para um conjunto de dados de referência gerados sob a hipótese nula, com o valor da estatística obtida do conjunto de dados a analisar. Assim, para se concluir sobre a significância do valor de um índice, será necessário determinar a distribuição do índice de validação, sob a hipótese nula de ausência de estrutura.

O uso de grafos como base teórica de suporte ao desenvolvimento dos índices definidos nos capítulos anteriores, facilita o estudo estatístico dos mesmos, permitindo a definição das suas distribuições sob modelos definidos no contexto da teoria de grafos aleatórios.

Assim, vão ser apresentadas as distribuições teóricas dos índices não normalizados e quando aplicados a partições não ajustadas.

7.1 Teoria de Grafos Aleatórios

A teoria de grafos aleatórios consiste na aplicação de métodos de probabilidades à teoria de grafos, tendo como objectivo estudar algumas propriedades importantes dos grafos.

7.1.1 Um pouco de História

A teoria de grafos aleatórios teve a sua origem numa série de oito artigos fundamentais publicados no período de 1959-1968 (ver [127]) por dois matemáticos húngaros, Paul Erdös e Alfred Rényi. A partir desta data, a teoria tem tido uma significativa evolução como um ramo independente da matemática discreta, localizado na intersecção da teoria de grafos com a teoria combinatória e probabilística. A teoria de grafos aleatórios tem uma vasta aplicabilidade em diversas áreas como ciência de computadores, redes de comunicações, ciências naturais e sociais e na própria matemática discreta.

Nos últimos anos, Béla Bollobás tornou-se o cientista líder na área de grafos aleatórios contribuindo com dezenas de artigos, culminando num livro editado em 1985 com uma segunda edição em 2001 [29]. Muitos outros investigadores têm trabalhado nesta área contribuindo para que a produção científica seja muito elevada, permitindo que a teoria de grafos aleatórios evolua rapidamente.

7.1.2 Modelos de Grafos Aleatórios

Um grafo aleatório é um grafo *construído* por algum *procedimento aleatório*, podendo ser usado no estudo probabilístico de algumas estruturas combinatórias [121].

O modelo introduzido por Erdös é muito natural e pode ser descrito como escolhendo um grafo aleatoriamente, com igual probabilidade, a partir do conjunto de cardinal 2^N sendo $N = \binom{n}{2}$ o número de todos os grafos com n vértices. Considere-se um espaço de probabilidades $(\mathbb{G}, \mathbb{F}, Prob)$, onde $\mathbb{G}$ é o conjunto de todos os grafos com n vértices, $\mathbb{F}$ é a família de todos os subconjuntos de $\mathbb{G}$ e para cada $g \in \mathbb{G}$, o valor de $Prob(g)$ dá-nos a probabilidade de se obter um grafo específico, g. A *construção* é representada por uma função do espaço de probabilidades numa adequada família de grafos [121].

Em geral, no contexto da teoria de grafos aleatórios, muitas propriedades são definidas assimptoticamente, quando o número de vértices tende para o infinito, sob o respectivo modelo.

7.1.2.1 Modelos $G_{n,p}$ e $G_{n,M}$

Os dois modelos mais populares na literatura em teoria de grafos aleatórios são o modelo $G_{n,p}$ e o modelo $G_{n,M}$ (designados em [121] por modelo binomial e modelo uniforme, respectivamente), constituindo uma plataforma de suporte ao desenvolvimento da maior parte dos resultados conhecidos nesta área.

O modelo $G_{n,p}$ considera todos os grafos de n vértices, nos quais os arcos são inseridos de modo independente e com probabilidade p, sendo esta a probabilidade de existir um arco entre dois vértices quaisquer. Por outras palavras, se G_0 é um grafo com n vértices e m arcos, então [29]

$$Prob(G_0) = Prob(G = G_0) = p^m q^{N-m} \tag{7.1}$$

sendo $q = 1 - p$ e $N = \left(\begin{smallmatrix} n \\ 2 \end{smallmatrix} \right)$ o número de todos os pares de vértices.

A maior parte da literatura de grafos aleatórios consideram o caso no qual $p \to 0$ e $n \to \infty$.

O modelo $G_{n,M}$ considera todos os grafos de n vértices tendo M arcos, no qual os grafos têm a mesma probabilidade de ocorrer. Verifica-se $0 \leq M \leq N$, $G_{n,M}$ tem $\left(\begin{smallmatrix} N \\ M \end{smallmatrix} \right)$ elementos e cada elemento ocorre com probabilidade $\left(\begin{smallmatrix} N \\ M \end{smallmatrix} \right)^{-1}$.

Estes dois modelos básicos são, em muitos casos, assimptoticamente equivalentes, se Np estiver perto de M ([121] e [29]). Seja A uma qualquer propriedade dos grafos aleatórios e seja G_1 um grafo aleatório gerado pelo modelo $G_{n,p}$ e G_2 um grafo aleatório gerado pelo modelo $G_{n,M}$. Considere-se $Prob(G_i \in A)$ para $i \in \{1, 2\}$, a probabilidade de o grafo G_i ter a propriedade A. Os dois grafos G_1 e G_2 são equivalentes para a propriedade A se $Prob(G_1 \in A) \to a$ para $n \to \infty$ é equivalente a $Prob(G_2 \in A) \to a$ para $n \to \infty$. Em [121] são estabelecidas algumas condições para as quais se verifica

esta equivalência.

7.1.2.2 Modelo $\tilde{G}_{[n_1,\ldots,n_k;p]}$

O modelo $\tilde{G}_{[n_1,\ldots,n_k;p]}$ representa um modelo para grafos aleatórios *k-partite*, com n elementos repartidos por k conjuntos independentes de $n_1,\ldots,n_k$ elementos, respectivamente; p é a probabilidade de existir um arco entre vértices de conjuntos diferentes.

Cada instância deste modelo tem k conjuntos independentes, sejam $C_1,\ldots,C_k$ de $n_1,\ldots,n_k$ vértices, respectivamente. Se $v \in C_i$ então o seu grau máximo (número máximo de arcos que ligam v a outros vértices) é $n - n_i$ uma vez que v só pode ser adjacente a vértices fora da classe à qual pertence. A soma dos graus máximos de todos os elementos da classe C_i é dado por $n_i(n - n_i)$. Percorrendo todas as classes obtemos o número máximo de arcos permitido num grafo *k-partite*:

$$\begin{aligned} E_{max} = \frac{1}{2}\sum_{i=1}^{k} n_i(n-n_i) &= \frac{1}{2}\left(\sum_{i=1}^{k} n_i n - \sum_{i=1}^{k} {n_i}^2\right) \\ &= \frac{1}{2}\left(n^2 - \sum_{i=1}^{k} n_i^2\right) \end{aligned} \tag{7.2}$$

Se G_0 é um grafo com n vértices e m arcos entre vértices de conjuntos independentes, então [29]

$$Prob(G_0) = Prob(G = G_0) = p^m q^{E_{max}-m} \tag{7.3}$$

7.1.3 Propriedades de Grafos Aleatórios

Seguindo o modelo $G_{n,p}$ da teoria de grafos aleatórios, os grafos aleatórios de n vértices podem ser considerados como entidades em evolução. Na fase inicial do processo os grafos não têm arcos, seguindo-se a sua inserção de modo aleatório, nas etapas seguintes. Uma das questões importantes é saber a partir de que etapa, o grafo aleatório apresenta determinada propriedade, por exemplo, a partir de que etapa o grafo passa a ser conexo. Um dos resultados mais interessantes em teoria de grafos

aleatórios indica que grande parte das propriedades dos grafos surge repentinamente. Ou seja, para um grafo de m arcos, a probabilidade de se verificar uma dada propriedade é muito baixa, enquanto que para $m+1$ arcos, a probabilidade de a mesma propriedade se verificar, já é muito alta [29].

A lista de publicações efectuadas no contexto da teoria de grafos aleatórios é vastíssima; no entanto, a maior parte dos estudos debruçam-se sobre um número reduzido de propriedades. Possivelmente um dos assuntos mais estudados relaciona-se com o número cromático ([29] e [121]), que consiste no menor número de cores necessárias e suficientes para colorir um grafo, de tal modo que dois vértices adjacentes não tenham a mesma cor. Grande parte dos resultados obtidos referem-se ao valor esperado do número cromático para grafos aleatórios ([23], [28], [153], [152], [3], [121], [29] e [5]).

Outros estudos debruçam-se sobre as propriedades dos graus dos vértices de um grafo aleatório. Alguns trabalhos estudam a sequência dos graus [29] de um grafo aleatório, tentando encontrar intervalos nos quais se localizam os graus dos vértices [25]. Outros estudos relacionam-se com a distribuição do número de vértices com um determinado grau num grafo simples ([184] e [27]) ou numa árvore [193]. A distribuição do grau máximo de uma grafo aleatório ([24]), ou dos graus dos vértices de um grafo [174], têm também suscitado algum interesse por parte dos investigadores.

Em teoria de grafos, dois vértices estão ligados se existir um caminho entre eles; define-se o comprimento desse caminho como o número de arcos nele contido, sendo a distância entre os vértices o comprimento mínimo de entre todos os caminhos que os ligam [41]. O diâmetro do grafo, que consiste na distância máxima entre pares de vértices de um grafo, é também alvo de alguma investigação quer nos grafos aleatórios simples [26], quer nos grafos aleatórios regulares, nos quais todos os vértices têm o mesmo grau [30], quer nos grafos bipartite (grafos *2-partite*) aleatórios [32].

Um subgrafo completo maximal de um grafo designa-se por *clique*, sendo o número de *clique* o número máximo de vértices de um *clique*. O número independente é a cardinalidade máxima de um conjunto de vértices independente (conjunto de vértices não adjacentes) [41]. Quer o número de *clique* ([31], [93] e [157]), quer o número independente ([25] e [79]) têm também sido alvo de algum estudo.

7.2 Modelos de Multigrafos Aleatórios

Nesta secção vão ser considerados os modelos para o estudo estatístico dos índices obtidos pelo método com atributos diferenciados, baseados em multigrafos. Considerem-se multigrafos com n vértices e com, no máximo, t arcos a ligar cada par de vértices.

7.2.1 Sobreposição Θ de Grafos Aleatórios

Na abordagem seguida por *Godehardt* em [86], os arcos são introduzidos um a um com probabilidade p, existindo, numa instância do modelo, no máximo $t\binom{n}{2}$ arcos. Na abordagem que se segue pretende-se que os arcos sejam inseridos por níveis, estando a cada nível associada uma probabilidade diferente. Ou seja, no nível 1 os arcos são inseridos com probabilidade p_1, no nível 2 os arcos são inseridos com probabilidade p_2, seguindo até ao nível t. Os n vértices e os arcos do nível 1 formam um grafo que corresponde a uma instância do modelo G_{n,p_1}, os mesmos vértices e os arcos do nível 2 formam um grafo que corresponde a uma instância do modelo G_{n,p_2} e assim sucessivamente. O multigrafo associado aos atributos diferenciados (AD) resulta de uma operação Θ aplicada aos t grafos G_{n,p_i}, para $i = 1, \ldots, t$.

Definição 7.1 *Define-se a operação Θ entre dois grafos, $G_1(V, E_1)$, $G_2(V, E_2)$, com n vértices e m_1 e m_2 arcos, respectivamente, da seguinte maneira*

$$G = G_1 \Theta G_2$$

em que $G = (V, E)$ é um multigrafo com n vértices tal que

$$\eta(v_i, v_j) = \eta^{G_1}(v_i, v_j) + \eta^{G_2}(v_i, v_j) \tag{7.4}$$

onde $E = E_1 \cup E_2$ e $\eta(v_i, v_j)$, $\eta^{G_1}(v_i, v_j)$ e $\eta^{G_2}(v_i, v_j)$ representam o número de arcos que unem os vértices v_i e v_j nos grafos G, G_1 e G_2, respectivamente.

Os grafos aleatórios representativos dos AD obtêm-se de:

$$G = \Theta_{i=1}^{t} G_i \tag{7.5}$$

sendo G_i o grafo aleatório ao nível i, gerado pelo modelo G_{n,p_i}, para $i = 1, \ldots, t$.

7.2.2 Modelo $\tilde{G}_{[n;p_1,\ldots,p_t]}$

O modelo $\tilde{G}_{[n;p_1,\ldots,p_t]}$ é o modelo $G_{n,p}$ estendido a atributos diferenciados, assim

$$\tilde{G}_{[n;p_1,\ldots,p_t]} = \Theta_{i=1}^{t} G_{n,p_i} \tag{7.6}$$

O modelo $\tilde{G}_{[n;p_1,\ldots,p_t]}$ representa um modelo de geração de multigrafos aleatórios de n vértices, existindo no máximo t arcos a unir dois vértices do multigrafo. Ao nível i, os arcos são distribuídos entre dois vértices com probabilidade p_i.

Seja $I = \{i : p_i \neq 0, i = 1, \ldots, t\}$ o conjunto dos índices que correspondem aos atributos presentes na classificação.

Definição 7.2 *Uma* I−projecção *de uma instância do modelo* $\tilde{G}_{[n;p_1,\ldots,p_t]}$ *no modelo* $G_{n,p}$*, origina uma instância deste modelo em que a probabilidade de existir um arco entre dois vértices é dada por:*

$$p = \prod_{i \in I} p_i$$

Admitindo que a ocorrência de arcos nos diferentes níveis são acontecimentos independentes.

7.2.3 Modelo $\tilde{G}_{[n_1,\ldots,n_k;p_1,\ldots,p_t]}$

O modelo $\tilde{G}_{[n_1,\ldots,n_k;p_1,\ldots,p_t]}$ representa um modelo de geração de multigrafos *k-partite* aleatórios, de n vértices e k conjuntos independentes de $n_1, \ldots, n_k$ elementos, respectivamente; p_i é a probabilidade de existir um arco entre vértices de conjuntos diferentes no nível i.

7.3 Distribuição do Número de Arcos

7.3.1 Modelo $G_{n,p}$

Sejam $N = \binom{n}{2}$ o número de pares não ordenados no conjunto de n vértices e p a probabilidade de existir um arco entre dois vértices quaisquer.

Seja M a v.a. representativa do número de arcos de um grafo aleatório do modelo $G_{n,p}$, então:

$$Prob(M = m) = \binom{N}{m} p^m (1-p)^{N-m},\ m = 0, 1, \ldots, N \tag{7.7}$$

é a probabilidade de um grafo de n vértices ter m arcos. A v.a. M segue então uma distribuição Binomial de parâmetros N e p,

$$M \sim Bi(N, p) \tag{7.8}$$

7.3.2 Modelo $\tilde{G}_{[n_1,\ldots,n_k;p]}$

Seja M a variável aleatória representativa do número de arcos numa instância do modelo $\tilde{G}_{[n_1,\ldots,n_k;p]}$. Então

$$M \sim Bi\left(E_{max}, p\right) \tag{7.9}$$

sendo p a probabilidade de existir um arco entre vértices de conjuntos independentes diferentes, no grafo aleatório *k-partite* e $E_{max} = \frac{1}{2}\left(n^2 - \sum_{i=1}^{k} n_i^2\right)$.

7.3.3 Modelo $\tilde{G}_{[n;p_1,\ldots,p_t]}$

Tal como foi visto na secção anterior, o multigrafo $\tilde{G}_{[n;p_1,\ldots,p_t]}$ resulta da operação *Sobreposição*, Θ, dos grafos G_{n,p_i} onde $i = 1, \ldots, t$ sendo t o número de blocos de atributos. Verificam-se as seguintes situações:

1. Em cada nível os $\begin{pmatrix} n \\ 2 \end{pmatrix}$ arcos são equiprováveis
2. Pode acontecer $p_i = p_j$ para alguns $i, j \in \{1, \ldots, t\}$ tais que $i \neq j$
3. Pode acontecer $p_i = 0$ para algum $i \in \{1, \ldots, t\}$
4. $0 \leq \sum_{i=1}^{t} p_i \leq t$

Seja M_i a v.a. representativa do número de arcos no grafo aleatório $\tilde{G}_{[n;p_1,\ldots,p_t]}$ no nível i, então, pela secção anterior,

$$M_i \sim Bi(N, p_i) \tag{7.10}$$

A probabilidade do multigrafo de n vértices ter m_1 arcos no primeiro nível, m_2 arcos no segundo nível, etc. m_t arcos no nível t, é dada por:

$$Prob(M_1 = m_1, M_2 = m_2, \ldots, M_t = m_t) = \prod_{i=1}^{t} \begin{pmatrix} N \\ m_i \end{pmatrix} p_i^{m_i} (1 - p_i)^{N - m_i} \tag{7.11}$$

que é o produto das probabilidades associadas às variáveis aleatórias binomiais independentes de parâmetros N e p_i cada uma.

Dada a relação entre as distribuições Binomial e Normal (Teorema de De Moivre-Laplace) e para N grande verifica-se

$$M_i \approx N\left(Np_i, \sqrt{Np_i(1 - p_i)}\right) \tag{7.12}$$

No modelo $\tilde{G}_{[n;p_1,\ldots,p_t]}$, tN é o número máximo de arcos num grafo aleatório; o número médio de arcos existente num instância do modelo é dado pelo número médio de arcos existente no nível 1 mais o número médio de arcos existente no nível 2, etc.

Associado ao modelo $\tilde{G}_{[n;p_1,\ldots,p_t]}$ considere-se M a v.a. representativa do número de arcos no grafo aleatório gerado pelo modelo, então

$$E(M) = \left[\sum_{i=1}^{t} p_i\right] \times N \tag{7.13}$$

representa o número médio de arcos numa instância do modelo $\tilde{G}_{[n;p_1,\ldots,p_t]}$.

7.3.4 Modelo $\tilde{G}_{[n_1,\ldots,n_k;p_1,\ldots,p_t]}$

Seja M_i a variável aleatória representativa do número de arcos, no nível i, de uma instância do modelo $\tilde{G}_{[n_1,\ldots,n_k;p_1,\ldots,p_t]}$, então

$$M_i \sim Bi(E_{max}, p_i) \tag{7.14}$$

sendo p_i a probabilidade de existir um arco no nível i e $E_{max} = \frac{1}{2}\left(n^2 - \sum_{\ell=1}^{k} n_\ell^2\right)$.

A probabilidade do multigrafo de n vértices ter m_1 arcos no primeiro nível, m_2 arcos no segundo nível, etc. m_t arcos no nível t, é dada por:

$$Prob(M_1 = m_1, M_2 = m_2, \ldots, M_t = m_t) = \prod_{i=1}^{t} \begin{pmatrix} E_{max} \\ m_i \end{pmatrix} p_i^{m_i}(1-p_i)^{E_{max}-m_i} \tag{7.15}$$

sendo o produto das probabilidades associadas às variáveis aleatórias binomiais independentes de parâmetros E_{max} e p_i no nível i.

Dada a relação entre as distribuições Binomial e Normal e para E_{max} grande verifica-se

$$M_i \approx N\left(E_{max}p_i, \sqrt{E_{max}p_i(1-p_i)}\right) \tag{7.16}$$

No modelo $\tilde{G}_{[n_1,\ldots,n_k;p_1,\ldots,p_t]}$, $t\sum_{\ell=1}^{k} n_\ell(n-n_\ell)$ é o número máximo de arcos num grafo aleatório; o número médio de arcos existente num instância do modelo é dado pelo número médio de arcos existente no nível 1 mais o número médio de arcos existente

no nível 2, etc.

Consideremos o modelo $\tilde{G}_{[n_1,\ldots,n_k;p_1,\ldots,p_t]}$ e seja $A(C_1, C_2, \ldots, C_k)$ o número de arcos no grafo *k-partite* aleatório gerado pelo modelo, então

$$E\{A(C_1, C_2, \ldots, C_k)\} = \left[\sum_{i=1}^{t} p_i\right] \times \left[\sum_{\ell=1}^{k} n_\ell(n - n_\ell)\right] \tag{7.17}$$

é o número médio de arcos numa instância do modelo $\tilde{G}_{[n_1,\ldots,n_k;p_1,\ldots,p_t]}$.

7.4 Distribuição do Índice de Classificabilidade

Pretende-se nesta secção estudar a distribuição do índice de classificabilidade IC sem normalização e aplicado ao grafo representativo dos dados sem que tenha sido alvo de qualquer processo de ajuste.

7.4.1 Modelo $\tilde{G}_{[n_1,\ldots,n_k;p]}$

Seja X a variável aleatória representativa do número de arcos em falta numa instância do modelo $\tilde{G}_{[n_1,\ldots,n_k;p]}$ para que seja *k-partite* completo, assim:

$$X = \frac{1}{2}\left(n^2 - \sum_{i=1}^{k} n_i^2 - 2M\right) \tag{7.18}$$

onde M é a v.a. representativa do número de arcos. Como já foi referido na Secção 7.3.2, $M \sim Bi(E_{max}, p)$ onde $E_{max} = \frac{1}{2}\left(n^2 - \sum_{i=1}^{k} n_i^2\right)$ e p é a probabilidade de existir um arco entre dois vértices pertencentes a conjuntos independentes diferentes.

A partir da distribuição que segue a v.a. M, podemos encontrar a distribuição que segue a v.a. X,

$$\begin{array}{l} M \sim Bi(E_{max}, p) \\ \quad \text{e} \qquad\qquad \Longrightarrow \quad X \sim Bi(E_{max}, 1-p) \\ X = E_{max} - M \end{array} \tag{7.19}$$

Assim, o valor de IC segue uma distribuição Binomial de parâmetros E_{max} e $q = 1-p$ no modelo $\tilde{G}_{[n_1,\ldots,n_k;p]}$.

7.4.2 Modelo $\tilde{G}_{[n_1,\ldots,n_k;p_1,\ldots,p_t]}$

Seja X_i a variável aleatória representativa do número de arcos em falta no nível i, de uma instância do modelo $\tilde{G}_{[n_1,\ldots,n_k;p_1,\ldots,p_t]}$, para que seja *k-partite* completo, assim

$$X_i \sim Bi(E_{max}, 1-p_i) \tag{7.20}$$

sendo p_i a probabilidade de existir um arco no nível i.

Considere-se $X_i = E_{max} - M_i$, assim a probabilidade do multigrafo de n vértices ter x_1 arcos em falta no primeiro nível, x_2 arcos em falta no segundo nível, etc. x_t arcos em falta no nível t, é dada por:

$$Prob(X_1 = x_1, X_2 = x_2, \ldots, X_t = x_t) = \prod_{i=1}^{t} \left(\begin{array}{c} E_{max} \\ x_i \end{array} \right) (1-p_i)^{x_i} p_i^{E_{max}-x_i} \tag{7.21}$$

sendo o produto das probabilidades associadas às variáveis aleatórias binomiais independentes de parâmetros E_{max} e $1-p_i$ no nível i.

Dada a relação entre as distribuições Binomial e Normal e para E_{max} grande verifica-se

$$X_i \approx N\left(E_{max}(1-p_i), \sqrt{E_{max}p_i(1-p_i)}\right) \tag{7.22}$$

No modelo $\tilde{G}_{[n_1,\ldots,n_k;p_1,\ldots,p_t]}$, $t\sum_{\ell=1}^{k} n_\ell(n-n_\ell)$ é o número máximo de arcos num grafo aleatório; o número médio de arcos é dado pelo número médio de arcos existente no nível 1 mais o número médio de arcos existente no nível 2, etc.

Considerando o modelo $\tilde{G}_{[n_1,\ldots,n_k;p_1,\ldots,p_t]}$ seja $A(C_1, C_2, \ldots, C_k)$ o número de arcos no grafo *k-partite* aleatório gerado pelo modelo, então

$$E\{A(C_1, C_2, \ldots, C_k)\} = \left[t - \sum_{i=1}^{t} p_i\right] \times \left[\sum_{\ell=1}^{k} n_\ell(n - n_\ell)\right] \tag{7.23}$$

representa o número médio de arcos em falta para que uma instância do modelo $\tilde{G}_{[n_1,\ldots,n_k;p_1,\ldots,p_t]}$ seja *k-partite*.

7.4.3 Modelo Uniforme

Nesta secção pretende-se encontrar a distribuição teórica do índice de classificabilidade não normalizado e para partições não ajustadas, no modelo uniforme.

7.4.3.1 Hipótese Nula

Pretende-se testar a hipótese nula de uniformidade a partir do valor observado do índice. Seja então

$$H_0 : \text{ "Ausência de classes"}$$

a hipótese nula de uniformidade ou de ausência de classes. Através dos testes de hipóteses pretendemos testar se temos forte evidência, para rejeitar ou não rejeitar a hipótese H_0 de uniformidade.

7.4.3.2 Distribuição do Índice

No modelo Uniforme consideram-se os indivíduos simulados como pontos uniformemente distribuídos por uma região A num espaço de dimensão t.

Recorde-se que na aplicação do método se verifica $\ell_{min} \leq \alpha \leq \ell_{max}$, sendo $\alpha = \ell_{min} + i \times h$ para $i = 1, \ldots, H$ e $h = \dfrac{\ell_{max} - \ell_{min}}{H}$. Para cada um destes valores do parâmetro de controlo α, o método é aplicado ao conjunto de dados, obtendo-se uma partição em k classes de n_i, $i = 1, \ldots, k$ elementos. Como já ficou bem patente nos capítulos anteriores, para cada valor de α é também obtido um valor do índice de classificabilidade. Este método selecciona a partição representativa dos dados aquela para a qual o valor do índice corresponde a uma diminuição relativamente a valores anteriores. No entanto, se o conjunto de dados não tiver uma estrutura em classes, o gráfico representativo dos valores de IC não permite identificar uma partição.

Pretende-se assim encontrar a distribuição do índice de classificabilidade sob a hipótese H_0, para cada α.

Seja D a v.a. representativa das dissemelhanças entre elementos do conjunto de dados. Admite-se que esta variável segue uma distribuição Normal [192] de parâmetros μ e σ. Como no modelo $G_{n,p}$, existe um arco com probabilidade p entre dois vértices, e no grafo representativo do conjunto de dados existe um arco $\{u, v\}$ se $d(u, v) \geq \alpha$, seja

$$p_\alpha = Prob(D \geq \alpha) \tag{7.24}$$

a probabilidade de existir um arco no grafo associado ao valor de α. Então

$$IC \sim Bi(E_{max}, 1 - p_\alpha) \tag{7.25}$$

onde

$$E_{max} = \frac{1}{2}\left(n^2 - \sum_{i=1}^{k} n_i^2\right) \tag{7.26}$$

sendo os cardinais n_i obtidos pela aplicação do método ao conjunto de dados a analisar. Dada a relação entre as distribuições Binomial e Normal e para E_{max} grande verifica-se

$$IC \approx N\left(E_{max}(1 - p_\alpha), \sqrt{E_{max} p_\alpha (1 - p_\alpha)}\right) \tag{7.27}$$

Consideremos um conjunto de dados com $n = 1000$ elementos distribuídos uniformemente na janela unitária. Determine-se a distribuição Normal empírica com os parâmetros $\hat{\mu}$ e $\hat{\sigma}$ estimados pelo método de máxima verosimilhança, ou seja,

$$\hat{\mu} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{ij} \tag{7.28}$$

e

$$\hat{\sigma}^2 = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (d_{ij} - \hat{\mu})^2 \tag{7.29}$$

sendo d_{ij} a distância euclideana entre os elementos x_i e x_j. Para o conjunto de dados considerado obteve-se: $\hat{\mu} = 0.7$ e $\hat{\sigma} = 0.2$. Verifica-se então:

$$\tilde{D} = \frac{D - \hat{\mu}}{\hat{\sigma}} \approx N(0,1) \tag{7.30}$$

e então

$$\begin{aligned} p_\alpha &= Prob(D \geq \alpha) \\ &= 1 - Prob(D < \alpha) \\ &= 1 - Prob\left(\frac{D - \hat{\mu}}{\hat{\sigma}} < \frac{\alpha - \hat{\mu}}{\hat{\sigma}}\right) \\ &\approx 1 - \Phi\left(\frac{\alpha - \hat{\mu}}{\hat{\sigma}}\right) \end{aligned} \tag{7.31}$$

7.4.3.3 Testes de Hipóteses

Nesta secção vai-se usar a distribuição do índice de classificabilidade não normalizado e não ajustado, para testar os valores dos índices obtidos para alguns conjuntos de dados analisados em capítulos anteriores, sob a hipótese nula H_0 de ausência de classes.

A tabela 7.1 indica os conjuntos de dados a testar. A tabela 7.2 representa os resultados obtidos. Para cada conjunto de dados, o valor de α escolhido é aquele que originaria a partição identificada pelo valor de $\tilde{IC}$, após a aplicação do ajuste de densidade. O valor IC_{Obs} indica o valor do índice obtido da aplicação do método não ajustado, sendo $IC'_{Obs} = \dfrac{IC_{Obs} - E_{max}(1 - p_\alpha)}{\sqrt{E_{max}(1 - p_\alpha)p_\alpha}}$, de acordo com a Equação (7.27) e onde p_α é dado pela Equação (7.31) e E_{max} é dado pela Equação (7.26).

Os valores obtidos permitem rejeitar a hipótese H_0 de uniformidade, ao nível de significância de 5%, considerando que para todos os conjuntos existe uma estrutura em classes.

Tabela 7.1: Conjuntos de dados testados.

Designação	k	Figura	Secção
FormaAlongada	4	4.3	4.5.1
PontosIsolados	6	4.5	4.5.2
FormaNaoConvexa	2	4.7	4.5.3
CincoClasses_1	5	4.11	4.6
QuatroClasses_1	4	4.12	4.7
RedeGrega	3	4.15	4.7
QuatroClasses_2	4	5.1	5.3.2
CincoClasses_2	5	5.2	5.3.2
CincoClassesNum	5	5.8	5.7

Tabela 7.2: Valores da Prova para os conjuntos de dados testados.

Designação	α	p_α	IC_{Obs}	E_{max}	IC'_{Obs}	$Prob(IC' \leq IC'_{Obs})$
FormaAlongada	0.803	0.302	116381	1777529	-1836.8	≈ 0
PontosIsolados	0.487	0.858	4108	416559	-244.3	≈ 0
FormaNaoConvexa	0.816	0.281	11826	112745	-458.7	≈ 0
CincoClasses_1	0.688	0.524	23085	437117	-560.2	≈ 0
QuatroClasses_1	0.830	0.258	9490	59925	-326.5	≈ 0
RedeGrega	0.487	0.855	408162	7309514	-684.5	≈ 0
QuatroClasses_2	0.905	0.154	13144	71542	-490.8	≈ 0
CincoClasses_2	0.682	0.534	17963	430677	-558.2	≈ 0
CincoClassesNum	0.551	0.773	75	97413	-168.6	≈ 0

7.5 Distribuição do Índice de Isolamento Absoluto

Nesta secção pretende-se apresentar a distribuição teórica do índice de isolamento absoluto, não normalizado e aplicado a uma partição não ajustada. Também o valor deste índice está associado a um determinado valor do parâmetro α.

7.5.1 Modelo $\tilde{G}_{[n_1,\ldots,n_k;p]}$

O grau de um vértice $v \in V$ representado por $d_G(v)$, segue uma distribuição Binomial de parâmetros $n-1$ e p sob o modelo $G_{n,p}$. Assim, seguindo o modelo $\tilde{G}_{[n_1,\ldots,n_k;p]}$ e fixo $v \in V$, seja X_v a v.a. que representa o grau de v, isto é X_v é v. a. tal que $X_v = d_G(v)$ para todos os vértices v pertencentes à classe C_i de n_i elementos. Então,

$$X_v \sim Bi(n - n_i, p), v \in C_i \tag{7.32}$$

ou seja, os graus dos vértices pertencentes à classe C_i seguem uma distribuição Binomial de parâmetros $n - n_i$ e p.

Seja

$$Y = \sum_{v \in C_i} X_v \tag{7.33}$$

a v.a. representativa da soma dos graus dos vértices de C_i em $\tilde{G}_{[n_1,\ldots,n_k;p]}$.

Admitindo que as v.a. X_v são independentes, vem

$$Y = \sum_{v \in C_i} X_v \sim Bi\left(\sum_{v \in C_i} (n - n_i), p\right) \tag{7.34}$$

ou seja

$$Y = \sum_{v \in C_i} X_v \sim Bi\left(n_i(n - n_i), p\right) \tag{7.35}$$

Assim, a soma das variáveis aleatórias X_v segue uma distribuição Binomial de parâmetros $n_i(n - n_i)$ e p.

Recorde-se a expressão de $\Delta^a{}_i$:

$$\Delta^a{}_i = n_i(n - n_i) - \sum_{v \in C_i} d_G(v) \tag{7.36}$$

se $|C_i| = n_i$.

Considerando X a v.a. representativa do valor de $\Delta^a{}_i$ e Y como definida em (7.33), vem

$$X = n_i(n - n_i) - Y \tag{7.37}$$

Como $Y \sim Bi\,(n_i(n - n_i), p)$ resulta de imediato que

$$X \sim Bi\,(n_i(n - n_i), 1 - p) \tag{7.38}$$

Dada a relação entre as distribuições Binomial e Normal e se $n_i(n - n_i)$ for grande verifica-se

$$X \approx N\left(n_i(n - n_i)(1 - p); \sqrt{n_i(n - n_i)(1 - p)p}\right) \tag{7.39}$$

7.5.2 Modelo Uniforme

No caso do modelo uniforme, verificam-se as mesmas distribuições com $p = p_\alpha = Prob(D \geq \alpha)$ onde D é uma v. a. representativa das dissemelhanças entre os elementos do conjunto de dados, que se admite seguir uma distribuição Normal.

7.5.2.1 Testes de Hipóteses

Nesta secção vai-se usar a distribuição do índice de isolamento absoluto não normalizado aplicado a uma partição não ajustada, para testar os valores dos índices obtidos

Tabela 7.3: Valores da Prova para o conjunto representado na figura 5.1.

Classe	α	p_α	IIA_{Obs}	$n_i(n-n_i)$	IIA'_{Obs}	$Prob(IIA' \leq IIA'_{Obs})$
C_1	1.354	0.0005	14654	38479	-5428.7	≈ 0
C_2	1.354	0.0005	775	13756	-4948.2	≈ 0
C_3	1.354	0.0005	604	8671	-3854.6	≈ 0
C_4	1.354	0.0005	32	6144	-3486.2	≈ 0
C_5	1.354	0.0005	5639	15664	-3580.3	≈ 0
C_6	1.354	0.0005	1109	10071	-3992.6	≈ 0
C_7	1.354	0.0005	9046	23100	-4132.9	≈ 0
C_8	1.354	0.0005	913	5031	-3595.55	≈ 0
C_9	1.354	0.0005	336	796	-728.7	≈ 0
C_10	1.354	0.0005	83	1584	-1686.2	≈ 0
C_11	1.354	0.0005	67	796	-1155.2	≈ 0

para os conjuntos de dados representados nas figuras 5.1 e 5.2 (Secção 5.3.2), sob a hipótese nula H_0 de ausência de classes.

Para cada conjunto de dados, o valor de α escolhido é aquele que identificou a partição representada nas respectivas figuras, após o ajuste, justificando-se assim a obtenção de um maior número de classes nas partições aqui utilizadas. Os resultados estão indicados nas tabelas 7.3 e 7.4. O valor IIA_{Obs} indica o valor do índice obtido da aplicação do método não ajustado, sendo $IIA'_{Obs} = \dfrac{IIA_{Obs} - n_i(n-n_i)(1-p_\alpha)}{\sqrt{n_i(n-n_i)p_\alpha(1-p_\alpha)}}$, de acordo com a Equação (7.39) e onde p_α é dado pela Equação (7.31), n_i é o número de elementos da classe C_i.

Os valores obtidos permitem rejeitar a hipótese H_0 de uniformidade, ao nível de significância de 5%, considerando os valores encontrados significativos.

7.6 Distribuição do Índice de Isolamento Relativo

Nesta secção pretende-se determinar a distribuição teórica do índice de isolamento relativo não normalizado e aplicado a uma partição não ajustada.

Tabela 7.4: Valores da Prova para o conjunto representado na figura 5.2.

Classe	α	p_α	IIA_{Obs}	$n_i(n-n_i)$	IIA'_{Obs}	$Prob(IIA' \leq IIA'_{Obs})$
C_1	0.682	0.5359	5072	69375	-206.5	≈ 0
C_2	0.682	0.5359	3445	141100	-331.2	≈ 0
C_3	0.682	0.5359	1593	163564	-368.5	≈ 0
C_4	0.682	0.5359	5500	79431	-223.2	≈ 0
C_5	0.682	0.5359	1484	165319	-371.1	≈ 0
C_6	0.682	0.5359	2379	39319	-160.5	≈ 0
C_7	0.682	0.5359	3855	51975	-178.3	≈ 0
C_8	0.682	0.5359	889	8919	-69.0	≈ 0
C_9	0.682	0.5359	3027	61644	-206.6	≈ 0
C_10	0.682	0.5359	891	5964	-48.74	≈ 0
C_11	0.682	0.5359	2464	24375	-113.65	≈ 0
C_12	0.682	0.5359	460	5964	-59.9	≈ 0
C_13	0.682	0.5359	196	2991	-43.71	≈ 0
C_14	0.682	0.5359	2061	21516	-108.3	≈ 0
C_15	0.682	0.5359	166	1996	-34.13	≈ 0
C_16	0.682	0.5359	1438	8919	-56.3	≈ 0
C_17	0.682	0.5359	151	1996	-34.77	≈ 0
C_18	0.682	0.5359	148	1996	-34.9	≈ 0
C_19	0.682	0.5359	304	1996	-27.9	≈ 0
C_20	0.682	0.5359	263	1996	-39.7	≈ 0
C_21	0.682	0.5359	140	999	-20.5	≈ 0

7.6.1 Modelo $\tilde{G}_{[n_1,\ldots,n_k;p]}$

Fixo $v \in C_i$, seja $X_{vj} = d_G^{C_j}(v)$ a v.a. representativa do grau de v restrito à classe C_j. Isto é, X_{vj} representa o número de arcos que unem o vértice v aos elementos de C_j (número de elementos de C_j que estão ligados ao vértice v).

Verifica-se:

$$X_{vj} \sim Bi(n_j, p) \tag{7.40}$$

Como se considera que os graus dos vértices dos elementos de C_i, restritos à classe C_j, são independentes entre si podemos escrever:

$$Y = \sum_{v \in C_i} X_{vj} \sim Bi(n_i n_j, p) \tag{7.41}$$

Recorde-se a definição do índice de isolamento relativo,

$$\Delta^r{}_i(C_j) = n_i\, n_j - \sum_{v \in C_i} d_G^{C_j}(v) \tag{7.42}$$

Considerando X a v.a. representativa do valor de $\Delta^r{}_i(C_j)$ e Y como definida em (7.41), vem

$$X = n_i n_j - Y \tag{7.43}$$

Como,

$$Y \sim Bi(n_i n_j, p) \tag{7.44}$$

resulta de imediato,

$$X \sim Bi(n_i n_j, 1-p) \tag{7.45}$$

Dada a relação entre as distribuições Binomial e Normal e se $n_i n_j$ for grande verifica-se

$$X \approx N\left(n_i n_j(1-p), \sqrt{n_i n_j p(1-p)}\right) \tag{7.46}$$

7.6.2 Modelo Uniforme

No caso do modelo uniforme, verificam-se as mesmas distribuições com $p_\alpha = Prob(D \geq \alpha)$ onde D é uma v. a. representativa das dissemelhanças entre os elementos do conjunto de dados, que se admite seguir uma distribuição Normal.

7.7 Dependência dos Arcos / Grafos de Markov

O modelo de grafos aleatórios usado na definição das distribuições teóricas pressupõe independência entre arcos, o que nem sempre se verifica. Alguns autores têm-se debruçado sobre este assunto e têm surgido na literatura alguns modelos em alternativa aos clássicos que tomam em consideração a possível dependência entre os arcos de um grafo aleatório. Por exemplo, no contexto dos modelos para redes sociais, em [174] e em [175] é apresentado um novo modelo, baseado em grafos aleatórios com uma qualquer distribuição dos graus dos vértices. O modelo verifica a propriedade: "a probabilidade de u e v estarem ligados por um arco é maior se eles tiverem um vizinho em comum", apresentando um índice que indica a probabilidade de dois vizinhos u e v de um dado vértice w, serem também vizinhos um do outro.

Grafos de Markov [71]

Os modelos estatísticos log-lineares podem ser usados para caracterizar grafos aleatórios com uma estrutura dependente geral, e com uma dependência de Markov, na qual só arcos incidentes podem ser condicionalmente dependentes.

Considere-se a sequência $Z = \{Z_1, \ldots, Z_m\}$ de m v.a. discretas e $M = \{1, \ldots, m\}$ o conjunto dos índices. Um *grafo de dependências* D sobre M define as dependências condicionais presentes entre os pares de v.a. em Z. O grafo D consiste no conjunto dos pares $\{i, j\} \in M \times M$ tais que Z_i e Z_j são condicionalmente dependentes das restantes variáveis do conjunto Z. Por exemplo, uma sequência $Z = \{Z_1, \ldots, Z_m\}$ de

v.a. independentes tem um grafo de dependências $D = \emptyset$.

Os grafos de dependências D sobre o conjunto dos índices M, pode ser usado para descrever estruturas dependentes de um qualquer conjunto de variáveis indexadas pelos elementos do conjunto M. Em particular, se se escolherem as v.a. como indicadores da presença ou ausência de arcos de um grafo aleatório G num conjunto de vértices N, então o conjunto dos índices destas variáveis aleatórias é o conjunto M de $\binom{n}{2}$ possíveis arcos de G. De facto, se G for um grafo aleatório em $N = \{1, \ldots, n\}$, os indicadores da presença ou ausência de arcos são dados pelos elementos da matriz adjacente de G. Sejam $Y = Y_{uv}$ para $\{u, v\} \in G$, essas $\binom{n}{2}$ variáveis binárias. Neste caso, o grafo de dependências D de G é um grafo não aleatório que especifica a estrutura de dependência entre as $\binom{n}{2}$ variáveis aleatórias Y_{uv}. Os vértices de D são os possíveis arcos de G, ou seja, os pares $\{u, v\} \in M$, e os arcos de D são os pares de arcos de G que são condicionalmente dependentes. Isto significa que D tem um arco entre $\{s, t\}$ e $\{u, v\}$ em M se e só se Y_{st} e Y_{uv} são condicionalmente dependentes das restantes variáveis Y.

Os autores em [71] apresentam a probabilidade de existir um grafo aleatório com um determinado grafo de dependências, que consiste num modelo log-linear com o número de parâmetros igual ao número de *cliques* possível no grafo de dependências. O número de parâmetros a calcular para encontrar a probabilidade referida, impossibilita a aplicação do método a grafos com um número de vértices acima de algumas dezenas, uma vez que para cada *clique* de tamanho r existem $2^r - 1$ subconjuntos não vazios que também são *cliques*. No caso dos grafos que obedeçam à propriedade de Markov, a qual define que D contém um arco entre os subconjuntos $\{s, t\}$ e $\{u, v\}$ em M, se eles forem disjuntos, significando que arcos não incidentes em G são condicionalmente independentes, os *cliques* são os triângulos e as estrelas.

7.8 Conclusão

Foi obtida neste capítulo a distribuição para o índice de classificabilidade não normalizado e não ajustado, sob a hipótese de que o conjunto das dissemelhanças entre os elementos do conjunto a classificar constitui uma amostra aleatória.

A determinação da distribuição do índice de avaliação de partições (IAP), requereria

pressupostos de independência entre o número de vizinhos comuns de pares de vértices, hipótese que nos parece demasiado forte.

Capítulo 8

Aplicações e Comparação com outros Métodos

Neste capítulo pretende-se avaliar o desempenho do Método de Partição baseado na Coloração de Grafos (MPCG) e do Índice de Classificabilidade Normalizado ($I\tilde{C}$), quando aplicados a conjuntos de dados reais e artificiais, comparando os resultados com os obtidos por outros métodos de classificação. Após uma breve descrição dos conjuntos de dados e dos métodos de classificação, são apresentados os resultados obtidos.

8.1 Conjuntos de Dados

Os conjuntos de dados reais e artificiais apresentados foram seleccionados com o objectivo de avaliar o desempenho dos métodos de classificação, perante conjuntos de dados com um elevado número de elementos, atributos (ou variáveis) mistos, pontos isolados e classes de formas arbitrárias.

Sempre que disponível é também apresentado o número de classes que a partição de referência indica existir no conjunto, assim como o resultado da aplicação do índice de validação τ_a (Secção 5.2) [98] a cada uma das partições de referência. O índice de avaliação de partições (IAP) não é calculado, uma vez que só é aplicável a partições obtidas pelo método MPCG.

Tabela 8.1: Conjuntos de dados reais. Dados de referência.

Conjunto Dados	n	Número de Atributos Numéricos	Número de Atributos Categóricos	k	τ_a
Cidades Gregas	5922	2	0	-	-
Iris	150	4	0	3	0.862
UNESCO	45	2	0	-	-
Zoo	101	1	15	7	0.912

8.1.1 Conjuntos de Dados Reais

São apresentados quatro conjuntos de dados reais. A tabela 8.1 indica o número de elementos, o número e o tipo de atributos de cada um dos conjunto de dados, assim como o número de classes k e o valor do índice τ_a, sempre que disponível.

O conjunto de dados *Cidades Gregas* representado na figura 8.1, foi disponibilizado pelo Prof. *Yannis Theodoridis* [1] e contém a localização de 5922 cidades gregas, representadas por pontos num espaço euclideano de duas dimensões. Para este conjunto de dados não existe partição de referência.

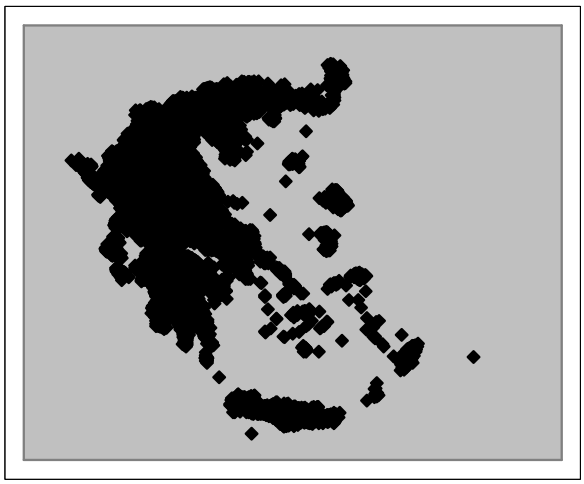

Figura 8.1: Conjunto constituído pela localização de 5922 cidades gregas.

Os conjunto de dados *Zoo*, *Iris* e UNESCO foram apresentados nas Secções 6.6.1, 6.6.2 e 6.6.3 respectivamente, no contexto dos atributos diferenciados. O conjunto *Zoo* é

[1] http://thalis.cs.unipi.gr/∼ytheod.

Tabela 8.2: Conjuntos de dados artificiais. Dados de referência.

Designação	k	τ_a	Tabela de Parâmetros do Gerador de Dados	Secção
C400_1	4	0.918	4.1 e 4.2	4.5.1
C400_2	6	0.929	4.3	4.5.2
C400_3	2	0.481	5.16	5.8
C500	2	0.731	4.4	4.5.3

constituído por 101 animais caracterizados por 16 atributos, sendo 15 binários e 1 numérico. Os atributos caracterizam os animais com o objectivo de os classificar em um de sete tipos de animais: mamíferos, pássaros, répteis, peixes, anfíbios, insectos e invertebrados. O conjunto *Iris* é constituído por 150 flores, 50 de cada uma das três variedades de Iris: Setosa, Versicolor e Virginica; caracterizadas por quatro atributos numéricos que descrevem a largura e o comprimento das pétalas e das sépalas de cada flor. O conjunto UNESCO é constituído por 45 países caracterizados por cinco atributos numéricos, indicando a percentagem de jovens entre os 11 e os 15 anos que frequentam o ensino primário e secundário (tabelas 6.19 e 6.20). Supõe-se que este conjunto tem duas classes: países participantes no programa WEI [2] e países pertencentes à OECD [3].

8.1.2 Conjuntos de Dados Artificiais

Quatro dos cinco conjuntos de dados artificiais considerados consistem em versões reduzidas (menor número de elementos) de alguns conjuntos de dados apresentados nos capítulos 4 e 5. A tabela 8.2 apresenta a designação dos conjuntos, o número de classes k e o valor do índice τ_a de cada uma das partições de referência. São ainda indicadas as tabelas de parâmetros do gerador de dados usados na geração dos conjuntos, e a secção onde os conjuntos de dados foram inicialmente apresentados. O quinto conjunto de dados artificiais, designado por GUE resultou da aplicação de um gerador de distâncias aleatórias ([97] e [96]). Usando os parâmetros do gerador de dados indicados na tabela 8.2 para cada um dos conjuntos, foram gerados 100 conjuntos de dados de cada tipo e as suas partições de referência.

[2] *World Education Indicators*

[3] *Organisation for Economic Co-operation and Development*

8.2 Métodos de Classificação

Nesta secção apresentam-se brevemente os métodos a serem comparados com os métodos MPCG. Quanto aos métodos MPCG, serão apresentados os resultados obtidos pelo método com ajuste de densidade (Secção 4.2.3) e com ajuste hierárquico (Secção 5.8).

8.2.1 Algoritmo METIS

No contexto dos algoritmos *graph partitioning* foi seleccionado o método METIS ([129], [130] e [128]) que tem como objectivo agrupar os vértices de um grafo em subconjuntos aproximadamente de igual tamanho, minimizando o número de arcos entre vértices de subconjuntos diferentes. Este método aceita como entrada um grafo de n vértices e m arcos, tendo como saída um mapeamento entre os vértices do grafo e as classes ou grupos de vértices.

Como, no nosso caso, o conjunto de arcos do grafo associado a cada conjunto de dados depende do valor do parâmetro de controlo α, e como o algoritmo METIS aceita como entrada o conjunto dos arcos, indirectamente este método necessita que se especifique o valor do parâmetro de controlo, para assim especificar o conjunto de arcos. Assim, para que os resultados deste algoritmo sejam comparáveis com os obtidos pelos métodos MPCG, o conjunto de arcos do grafo de entrada para o algoritmo METIS, corresponde ao grafo definido para o valor de α correspondente à partição identificada pelo método MPCG com ajuste de densidade (que foi única em todos os casos considerados). No entanto, no grafo de entrada do método METIS, existe um arco entre u e v se $dist(u, v) < \alpha$, uma vez que este método tenta minimizar o número de arcos entre subconjuntos diferentes; ao contrário do método MPCG, que tenta maximizar esse número, uma vez que no grafo representativo dos dados se verifica a existência de um arco se $dist(u, v) \geq \alpha$.

8.2.2 Métodos Não-Hierárquicos da Família VL

Os métodos da família VL (Secção 2.1.3) de classificação não hierárquica usados baseiam-se na utilização da função de distribuição de uma estatística das semelhanças VL, como critério de agregação de um elemento a uma determinada classe. Os métodos usados são os seguintes: NHMEAN que se baseia na estatística da média, NHVMED que se baseia na estatística mediana, NHVMG que usa a média geométrica e NHVC

que se baseia no centro amostral ([36], [181], [210] e [37]).

8.2.3 k-Prototype

O algoritmo *k-prototype* [111] é uma extensão do algoritmo das *k-médias* suportando atributos mistos (numéricos e/ou categóricos). Neste algoritmo a dissemelhança entre dois elementos foi definida na Secção 3.9 pela Equação 3.20, e é dada pelo quadrado da distância euclideana entre os atributos numéricos, e o número de categorias não coincidentes para cada um dos atributos categóricos.

8.3 Resultados

Nesta secção serão apresentados os resultados da aplicação dos métodos seleccionados aos conjuntos de dados reais e artificiais.

8.3.1 Conjuntos de Dados Reais

Uma vez que os conjuntos de dados reais *Iris*, UNESCO e *Zoo* já foram analisados no capítulo 6, só serão apresentados os resultados obtidos.

8.3.1.1 Conjunto de Dados Cidades Gregas

Para este conjunto de dados só foram aplicados os métodos MPCG e o algoritmo *k-prototype* uma vez que os outros métodos não suportam conjuntos de dados com tão elevado número de elementos.

Para este exemplo foi considerado um único bloco com os dois atributos e como função global a distância euclideana.

A figura 8.2 representa o gráfico obtido pelo método MPCG com ajuste de densidade para este conjunto de dados. A parte superior mostra o resultado da aplicação do método para $\ell_{min} = d_{min}$, $\ell_{max} = d_{max}$ e $H = 50$, sendo d_{min} e d_{max} e distância euclideana mínima e máxima, respectivamente, entre todos os pares de elementos do conjunto. Como se pode constatar da figura, os valores correspondentes à partição

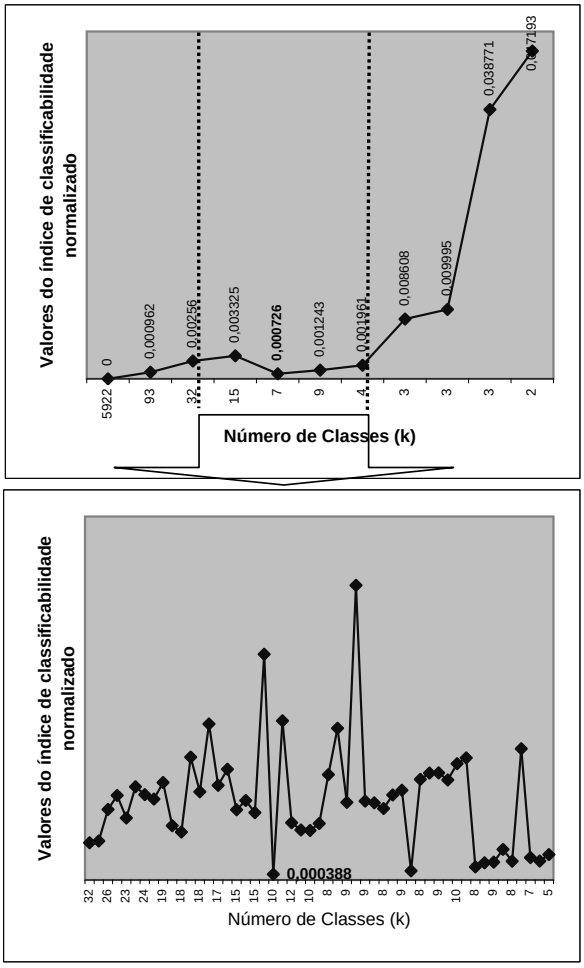

Figura 8.2: Gráfico $\tilde{IC}/k$ obtido pelo ajuste de densidade para o conjunto de dados cidades gregas.

escolhida pelo método são $\alpha = 79153.82$, $\tilde{IC} = 0.000726$ e $k = 7$. Como o valor do índice de classificabilidade não é zero, nem corresponde a uma descida acentuada em relação ao valor anterior do índice, o método foi novamente aplicado com os valores $\ell_{min} = 39793.32$ (correspondendo a $k = 32$), $\ell_{max} = 118514.3$ (correspondendo a $k = 4$) e $H = 50$. Para este caso foi obtida a partição representada na figura 8.3 e com os valores associados de $\alpha = 66558,47$, $\tilde{IC} = 0.000388$ e $k = 10$.

No caso da aplicação do método MPCG com ajuste hierárquico considerou-se $k = 15$, valor este estabelecido após uma análise da partição em 10 classes resultante da aplicação do método com ajuste de densidade, na qual algumas classes que deveriam ser consideradas separadas foram reunidas.

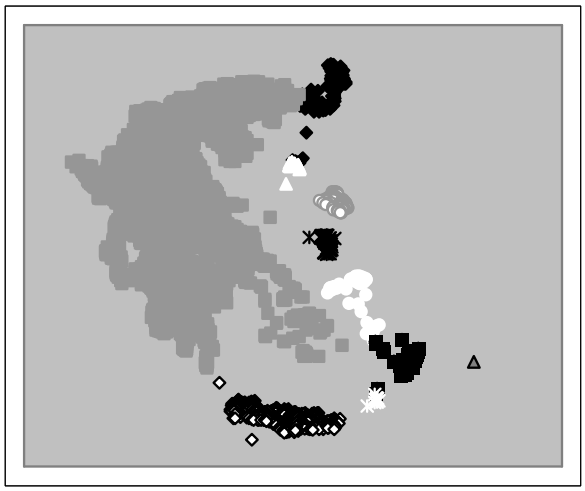

Figura 8.3: Partição em dez classes do conjunto cidades gregas, identificada pelo método com ajuste de densidade.

Tabela 8.3: Resultados para o Conjunto *Cidades Gregas*.

Método	k	α	$\tilde{IC}$	Figura
MPCG/Ajuste de Densidade	10	66558.5	0.000388	(8.3)
MPCG/Ajuste Hierárquico (k=15)	15	39793.3	0.000007	(8.5)
k-Prototype (k = 15)	15	-	-	(8.6)

Assim, para o ajuste hierárquico obteve-se o gráfico representado na figura 8.4 com os valores associados de $\alpha = 39793.32$, $\tilde{IC} = 0.000007$ e $k = 15$. Note-se que o valor do índice de classificabilidade neste caso é menor que para o caso anterior, correspondendo assim a uma partição "melhor", o que pode ser constatado na figura 8.5. O elevado número de elementos impossibilitou a aplicação do ajuste hierárquico com k desconhecido.

Quanto à aplicação do método *k-prototype*, pelas razões já apontadas também foi considerado $k = 15$, e resultou a distribuição em classes representada na figura 8.6. Esta partição é bastante diferente das obtidas pelos métodos anteriores, resultante do facto de este método, sendo baseado no método das *k-médias*, tender a favorecer a formação de classes esféricas.

A tabela 8.3 apresenta um resumo dos resultados obtidos.

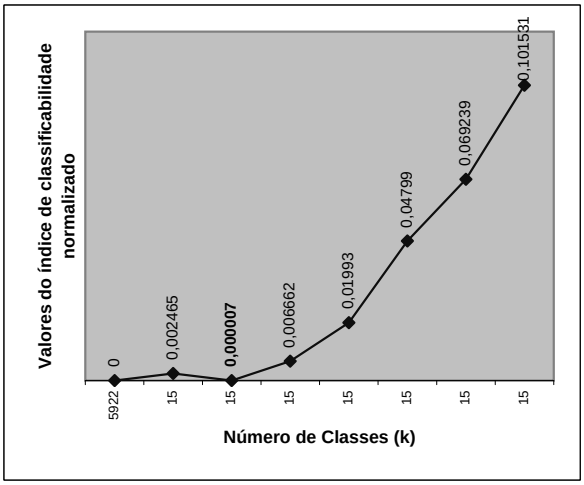

Figura 8.4: Gráfico $\tilde{IC}/k$ obtido pelo ajuste hierárquico ($k = 15$) para o conjunto de dados cidades gregas.

8.3.1.2 Conjunto de Dados *Iris*

Os diferentes métodos MPCG foram aplicados ao conjunto de dados *Iris* na Secção 6.6.2, estando os resultados representados na tabela 8.4.

O grafo associado ao valor de α identificado pelo método MPCG com atributos não diferenciados e ajuste de densidade, foi dado como entrada para o método METIS do qual se obtive o resultado representado na tabela 8.4. Como se pode ver na tabela, para a partição obtida por este método, o índice de *Rand* Corrigido de *Hubert* e *Arabie* (IRC) [113] mostrou-se inferior ao obtido pelo método das partições φ-projectadas ($\varphi = 4$) e partição consenso *medoid*. No entanto, esse valor é superior aos obtidos para todos os métodos restantes.

Para os métodos VL obtiveram-se valores do índice IRC inferiores ao do método METIS e inferiores aos dos métodos MPCG de partições φ-projectadas ($\varphi = 4$) e partição consenso *medoid*, apresentando valores superiores aos correspondentes a todos os outros métodos. Quanto aos valores do índice τ_a, estes métodos apresentaram valores semelhantes aos obtidos pelo método METIS, mas inferiores aos obtidos para os métodos MPCG, excepto para o caso da partição consenso *medoid*.

No caso do algoritmo *k-prototype* para $k = 3$ obteve-se uma partição em três classes, no entanto, com valores, quer do índice de validação τ_a, quer do índice IRC, inferiores

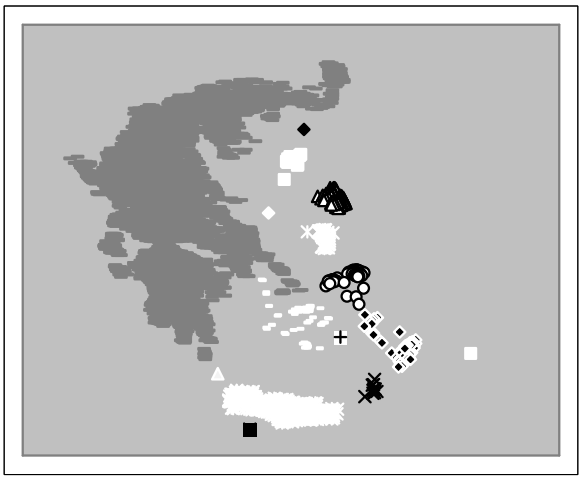

Figura 8.5: Partição em quinze classes do conjunto cidades gregas, obtida por ajuste hierárquico com $k = 15$.

aos registados para todos os outros casos testados, não sendo nenhuma das classes identificada correctamente.

O melhor resultado foi obtido com o método MPCG para partições φ-projectadas ($\varphi = 4$ e $t_b = 4$), verificando-se $IRC = 0.859$.

8.3.1.3 Conjunto de Dados *UNESCO*

Os diferentes métodos MPCG foram aplicados ao conjunto de dados UNESCO na Secção 6.6.3, estando os resultados obtidos representados na tabela 8.5. Como não existe partição de referência não foram apresentados os valores do índice IRC. Para os métodos que têm como parâmetro de entrada o número de classes, considerou-se $k = 2$, uma vez que os dados apresentam dois conjuntos de países: países da OECD e países participantes no programa WEI. O método para partições φ-projectadas não foi aplicado porque o conjunto de dados tem cinco atributos, tornando-se computacionalmente muito pesado obter um resultado.

Para o caso do método MPCG com ajuste hierárquico, obteve-se um valor do índice τ_a mais elevado, no entanto, a partição identificada é constituída por duas classes em que uma delas tem o país Zimbabwe, estando todos os restantes países noutra classe.

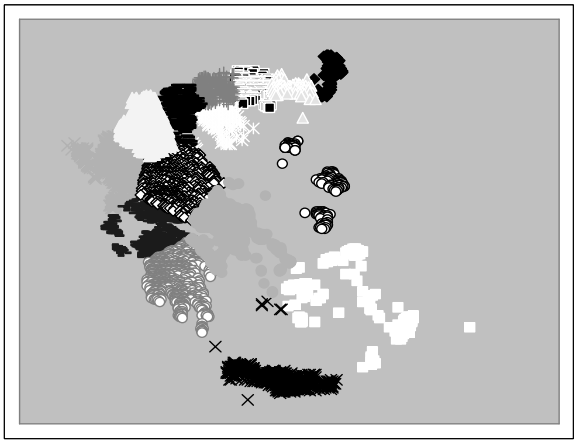

Figura 8.6: Partição em quinze classes do conjunto cidades gregas, obtida pelo método *k-prototype*.

Para o método METIS, obteve-se o pior valor do índice τ_a. O resultado justifica-se pela facto de este método tentar encontrar classes com um número de elementos aproximadamente igual, o que não é o caso do conjunto UNESCO, tendo uma classe 15 elementos e a outra 30 elementos. No entanto, nas duas partições obtidas, o método reuniu todos os países participantes no programa WEI.

No caso dos métodos VL, foi obtida a mesma partição para os quatro métodos, correspondente ao valor do índice τ_a superior aos obtidos pelo método METIS mas inferior relativamente aos outros métodos. Esta partição é constituída por duas classes contendo uma delas todas os países OECD, excepto o México e a Turquia; nesta classe foram ainda incluídos os seguintes países participantes no programa WEI: Argentina, Peru e Uruguai. Todos os restantes países pertencem à outra classe.

No caso do método *k-prototype* foi identificada uma partição que apresenta um valor de τ_a superior aos encontrados com o método METIS e com os métodos VL, no entanto, tendo em conta as tabelas 6.19 e 6.20, esta não parece uma partição de melhor qualidade. O método agrupou numa classe os países: Egipto, Indonésia, Jordânia, Paraguai, Filipinas, Tunísia, Zimbabwe, México e Turquia, ficando os restantes reunidos numa segunda classe.

Tabela 8.4: Resultados para o Conjunto *Iris*.

Método	k	τ_a	IRC
MPCG / Ajuste de Densidade	2	0.929	0.568
MPCG / Ajuste Hierárquico (k = 3)	3	0.930	0.564
MPCG / Ajuste Hierárquico (k desconhecido)	2	0.929	0.568
MPCG / Partição φ-projectada / $\varphi = 4$ / $t_b = 4$	5	0.858	0.859
MPCG / Partição φ-projectada / $\varphi = 1$ / $t_b = 2$	2	0.929	0.568
MPCG / Partição Consenso Mediana / $t_b = 4$	3	0.931	0.560
MPCG / Partição Consenso Mediana / $t_b = 2$	2	0.929	0.568
MPCG / Partição Consenso *Medoid* / $t_b = 4$	4	0.850	0.837
MPCG / Partição Consenso *Medoid* / $t_b = 2$	2	0.929	0.568
METIS ($\alpha = 1.218$, m = 3387, k = 3)	3	0.877	0.786
Métodos VL / NHVMEAN (k = 3)	3	0.856	0.635
Métodos VL / NHVMED (k = 3)	3	0.856	0.635
Métodos VL / NHVMG (k = 3)	3	0.853	0.624
Métodos VL / NHVC (k = 3)	3	0.853	0.635
k-Prototype (k = 3)	3	0.568	0.154

Tabela 8.5: Resultados para o Conjunto *UNESCO*.

Método	k	τ_a
MPCG / Ajuste de Densidade	7	0.939
MPCG / Ajuste Hierárquico (k = 2)	2	0.945
MPCG / Ajuste Hierárquico (k desconhecido)	2	0.945
MPCG / Partição Consenso Mediana	9	0.909
MPCG / Partição Consenso *Medoid*	6	0.931
METIS ($\alpha = 21.246$, m = 453, k = 2)	2	0.685
Métodos VL / NHMEAN (k = 2)	2	0.783
Métodos VL / NHVMED (k = 2)	2	0.783
Métodos VL / NHVMG (k = 2)	2	0.783
Métodos VL / NHVC (k = 2)	2	0.783
k-Prototype / (k = 2)	2	0.887

Tabela 8.6: Resultados para o Conjunto *Zoo*.

Método	k	τ_a	IRC
MPCG / Ajuste de Densidade	19	0.992	0.729
MPCG / Ajuste Hierárquico (k = 7)	7	0.945	0.808
MPCG / Ajuste Hierárquico (k desconhecido)	15	0.989	0.761
MPCG / Partição φ-projectada / $\varphi = 1$ / $t_b = 2$	7	0.923	0.517
MPCG / Partição Consenso Mediana / $t_b = 16$	11	0.827	0.887
MPCG / Partição Consenso Mediana / $t_b = 2$	21	0.988	0.715
MPCG / Partição Consenso *Medoid* / $t_b = 16$	2	0.707	0.585
MPCG / Partição Consenso *Medoid* / $t_b = 2$	10	0.922	0.807
METIS ($\alpha = 2.073757$, m = 276, k = 7)	7	0.899	0.381
k-Prototype (k = 7)	7	0.861	0.426

8.3.1.4 Conjunto de Dados *Zoo*

Este conjunto de dados é descrito por 16 atributos, sendo 15 deles binários. A função de dissemelhança usada foi a função global definida na Secção 3.9 pela Equação 3.20.

Para este conjunto de dados não foram aplicados os métodos VL não hierárquicos, porque não aceitam atributos mistos.

A tabela 8.6 representa os resultados obtidos pela aplicação dos métodos.

No caso do método METIS, como se pode ver na tabela 8.6, para o grafo correspondente ao valor de α obtido pelo método MPCG com ajuste de densidade com atributos não diferenciados, obtive-se uma partição de qualidade inferior, quer a nível do índice de validação $\tau_a = 0.899$, quer a nível de semelhança com a partição de referência $IRC = 0.381$. Tal como no caso do conjunto UNESCO, também para este caso a menor semelhança da partição obtida pelo método METIS com a partição de referência, se justifica pelo facto das classes terem disparidade de tamanhos.

A eficácia do método *k-prototype* foi pior que a dos métodos MPCG, mas melhor que a do método METIS, relativamente à semelhança com a partição de referência.

8.3.2 Conjuntos de Dados Artificiais

Para os conjuntos de dados artificiais considerados, não foram aplicados os métodos de atributos diferenciados, uma vez que não se justifica a sua aplicação a dados caracterizados com dois atributos numéricos. No caso do conjunto GUE, não é possível aplicar aqueles métodos uma vez que só está disponível a matriz de dissemelhança.

Usando os parâmetros do gerador de dados indicados na tabela 8.2 para cada um dos conjuntos, foram gerados 100 conjuntos de dados de cada tipo e as suas partições de referência. O método foi aplicado a cada conjunto de dados e a partição seleccionada foi comparada com a partição de referência, usando o índice IRC.

A tabela 8.7 apresenta a média e o desvio padrão dos valores do índice IRC entre as partições identificadas por cada um dos métodos, e as partições de referência dos 100 conjuntos de dados artificiais do tipo C400_1, constituídos por três classes de forma alongada e uma de forma esférica (figura 4.3). No caso do método METIS, para este exemplo e para os que se seguem, considerou-se o valor de α associado à partição identificada pelo método MPCG com ajuste de densidade. Como se pode ver pela tabela, os melhores resultados foram obtidos para os métodos MPCG com ajuste hierárquico, sendo os piores obtidos pelos métodos VL. Os métodos METIS e *k-prototype* apresentaram resultados semelhantes. Finalmente, note-se que os resultados obtidos para os métodos MPCG foram um pouco inferiores aos obtidos na Secção 5.8 e representados na tabela 5.13. O facto é que a versão do conjunto de dados com 400 elementos usada neste capítulo (recorde-se que a versão estudada anteriormente tem 2000 elementos), em alguns conjuntos, originou agrupamentos nas classes, nomeadamente nas elípticas, que foram identificadas como classes isoladas pelos métodos.

A tabela 8.8 apresenta a média e o desvio padrão dos valores do índice IRC entre as partições identificadas por cada um dos métodos, e as partições de referência dos 100 conjuntos de dados artificiais do tipo C400_2, constituído por 400 elementos, sendo 80% distribuídos por cinco classes e 20% distribuídos uniformemente pela janela que contém os restantes elementos (conjuntos semelhantes ao representado na figura 4.5, tendo, no entanto, uma maior percentagem de pontos isolados). A partição de referência é constituída por seis classes, agrupando-se numa delas todos os pontos isolados. Como se pode ver pela tabela os melhores resultados foram obtidos para os métodos MPCG com ajuste de densidade e hierárquico para k variável, sendo os piores obtidos pelos

Tabela 8.7: Média e desvio padrão dos valores do índice IRC de comparação com a partição de referência, para 100 conjuntos de dados do tipo C400_1.

Método	Média	Desvio Padrão
MPCG / Ajuste de Densidade	0.858	0.09
MPCG / Ajuste Hierárquico (k = 4)	0.986	0.054
MPCG / Ajuste Hierárquico (k desconhecido)	0.902	0.091
METIS (k = 4)	0.894	0.103
Métodos VL / NHMEAN (k = 4)	0.591	0.113
Métodos VL / NHVMG (k = 4)	0.605	0.114
Métodos VL / NHVC (k = 4)	0.580	0.124
k-Prototype (k = 4)	0.862	0.150

métodos VL. Os métodos METIS e *k-prototype* apresentaram resultados semelhantes, sendo aquele o mais eficaz. Finalmente, note-se que os resultados obtidos para os métodos MPCG foram inferiores aos obtidos na Secção 5.8 e representados na tabela 5.14. O facto é que a versão do conjunto de dados com 400 elementos usada neste capítulo (recorde-se que a versão estudada anteriormente tem 1000 elementos), tem 20% de pontos isolados enquanto a versão anterior tinha 10%. Como o conjunto tem um menor número de elementos, e maior percentagem de pontos isolados, estes mascaram mais as classes. No caso do ajuste hierárquico, a diminuição de eficácia foi mais evidente uma vez que este método tem uma certa tendência para agrupar classes não totalmente separadas. Em geral, o que se verificou é que os métodos por ajuste hierárquico identificaram as cinco classes, no entanto, cada uma das classes incluiu um maior número de pontos isolados. No entanto, saliente-se que apesar desta diminuição de eficácia justificada, o método MPCG com ajuste de densidade foi o mais eficaz de entre os métodos comparados.

A tabela 8.9 apresenta a média e o desvio padrão dos valores do índice IRC entre as partições identificadas por cada um dos métodos, e as partições de referência dos 100 conjuntos de dados artificiais do tipo C400_3, constituído por duas classes, sendo uma esférica e outra constituída por um anel que circunda a classe anterior (figura 5.9). Como se pode ver pela tabela, os melhores resultados foram obtidos para os métodos MPCG com ajuste hierárquico, realçando a eficácia destes métodos na identificação de classes de formas não esféricas. O funcionamento base do método *k-prototype*, baseado no método das *k-médias*, justifica o resultado obtido para aquele método. O método METIS mostrou ser eficiente na identificação das classes não esféricas, tendo

Tabela 8.8: Média e desvio padrão dos valores do índice IRC de comparação com a partição de referência, para 100 conjuntos de dados do tipo C400_2.

Método	Média	Desvio Padrão
MPCG / Ajuste de Densidade	0.754	0.090
MPCG / Ajuste Hierárquico (k = 5)	0.594	0.144
MPCG / Ajuste Hierárquico (k desconhecido)	0.710	0.085
METIS (k = 5)	0.688	0.012
Métodos VL / NHMEAN (k = 5)	0.452	0.046
Métodos VL / NHVMG (k = 5)	0.425	0.005
Métodos VL / NHVC (k = 5)	0.354	0.001
k-Prototype (k = 5)	0.628	0.082

Tabela 8.9: Média e desvio padrão dos valores do índice IRC de comparação com a partição de referência, para 100 conjuntos de dados do tipo C400_3.

Método	Média	Desvio Padrão
MPCG / Ajuste de Densidade	0.611	0.128
MPCG / Ajuste Hierárquico (k = 2)	0.996	0.021
MPCG / Ajuste Hierárquico (k desconhecido)	0.969	0.054
METIS (k = 2)	0.846	0.029
k-Prototype (k = 2)	0.584	0.030

sensivelmente o mesmo número de elementos.

A tabela 8.10 apresenta a média e o desvio padrão dos valores do índice IRC entre as partições identificadas por cada um dos métodos, e as partições de referência dos 100 conjuntos de dados artificiais do tipo C500, constituído por duas classes de formas não convexas (figura 4.7). Como se pode ver pela tabela, todos os métodos, excepto o *k-prototype*, mostraram-se razoavelmente eficázes na identificação das classes, sendo os mais eficázes os métodos MPCG com ajuste hierárquico e o método METIS. Mais uma vez o método *k-prototype* apresentou o pior desempenho.

A tabela 8.11 apresenta a média e o desvio padrão dos valores do índice τ_a das partições identificadas por cada um dos métodos, dos 100 conjuntos de dados artificiais do tipo GUE. Este conjunto é constituído por 200 elementos caracterizados por 10 atributos numéricos, distribuídos por cinco classes esféricas, resultado da geração de distâncias

Tabela 8.10: Média e desvio padrão dos valores do índice IRC de comparação com a partição de referência, para 100 conjuntos de dados do tipo C500.

Método	Média	Desvio Padrão
MPCG / Ajuste de Densidade	0.942	0.038
MPCG / Ajuste Hierárquico (k = 2)	1	0
MPCG / Ajuste Hierárquico (k desconhecido)	1	0
METIS (k = 2)	0.960	0.028
Métodos VL / NHMEAN (k = 2)	0.846	0.023
Métodos VL / NHVMG (k = 2)	0.844	0.024
Métodos VL / NHVC (k = 2)	0.836	0.065
k-Prototype (k = 2)	0.525	0.001

Tabela 8.11: Média e desvio padrão dos valores do índice τ_a das partições dos 100 conjuntos de dados do tipo GUE.

Método	Média	Desvio Padrão
MPCG / Ajuste de Densidade	0.967	0.006
MPCG / Ajuste Hierárquico (k = 5)	0.937	0.064
MPCG / Ajuste Hierárquico (k desconhecido)	0.937	0.064
METIS (k = 5)	0.947	0.029

euclideanas aleatórias. Como para estes conjuntos de dados não existe partição de referência, usou-se o índice τ_a em vez do índice IRC. O método *k-prototype* não foi usado uma vez que este não lê matrizes de dissemelhança. A eficácia dos métodos testados foi boa.

8.4 Conclusão

Neste capítulo os métodos apresentados em capítulos anteriores foram aplicados a quatro conjuntos de dados reais (Cidades Gregas, Iris, UNESCO e Zoo) e a 100 execuções de cada um dos cinco conjuntos de dados artificiais (C400_1, C400_2, C400_3, C500 e GUE), comparando o resultado com seis métodos de classificação (METIS, NHMEAN, NHVMED, NHVMG, NHVC e *k-prototype*).

No caso do conjunto de dados Cidades Gregas, os métodos MPCG só foram com-

parados com o método *k-prototype* devido ao elevado número de elementos (5922). Para este conjunto de dados, o método MPCG com ajuste hierárquico apresentou um melhor resultado uma vez que o conjunto de dados é composto por classes de formas arbitrárias. O método *k-prototype* não identificou eficazmente uma partição para o conjunto de dados, uma vez que este é baseado no método das *k-médias*, não muito eficaz na detecção de classes de forma não esférica.

No caso do conjunto de dados *Iris* o método que apresentou um pior desempenho foi o *k-prototype* apresentando um valor $IRC = 0.154$ com a partição de referência, apresentando também o valor mais baixo do índice τ_a. Pelos resultados obtidos verifica-se que o índice τ_a identifica como melhores partições as obtidas pelos métodos MPCG com ajuste hierárquico para k fixo e partição consenso mediana, constituídas por três classes, em que uma das classes tem um número reduzido de elementos. O melhor resultado foi obtido para o método das partições φ-projectadas para $\varphi = 4$ e $t_b = 4$ com $IRC = 0.859$.

Para o conjunto de dados da UNESCO os melhores resultados relativamente ao índice τ_a foram obtidos para os métodos MPCG e os piores resultados foram obtidos para o método METIS.

Para o conjunto de dados com atributos mistos *Zoo*, a partição mais semelhante com a partição de referência foi a obtida para o método MPCG com ajuste hierárquico com k fixo, obtendo-se um valor $IRC = 0.808$. O pior desempenho foi apresentado pelo método METIS apresentando um valor $IRC = 0.381$.

Em relação aos conjuntos de dados artificiais, mais um vez se mostrou que o método MPCG com ajuste hierárquico é muito eficaz na detecção de classes de forma arbitrária sendo o ajuste de densidade mais robusto quanto à presença de pontos isolados. O método METIS é muito eficaz na identificação de classes de qualquer formato, desde que as classes tenham aproximadamente o mesmo número de elementos. O método *k-prototype* mostrou um desempenho inferior na identificação de classes de forma não esférica. Relativamente aos métodos VL, estes mostraram um bom desempenho na identificação de classes de forma não esférica, sofrendo um decréscimo acentuado de eficácia quando aplicados a conjuntos de dados com pontos isolados.

Concluindo, podemos dizer que de uma forma geral, a eficácia dos métodos MPCG em conjuntos de dados com elevado número de elementos, atributos mistos, pontos isolados e classes de formas arbitrárias, foi superior à dos métodos com os quais foram comparados. Realça-se ainda que o único método que não requer qualquer parâmetro de entrada é o método MPCG com ajuste de densidade. Apesar disso, este método mostrou-se eficiente na detecção de classes de formas esféricas, é robusto na presença de pontos isolados e apresenta uma eficácia razoável na detecção de classes de formas não esféricas. Os métodos os ajuste hierárquico são muito eficientes na detecção e classes de forma não esférica.

Capítulo 9

Conclusão e Trabalho Futuro

A presente dissertação tentou dar resposta a alguns problemas importantes em análise classificatória, destacando-se, pela sua importância, a determinação do número de classes do conjunto de dados, a identificação de conjuntos de dados sem uma estrutura em classes e a avaliação da qualidade dos resultados de uma classificação. Estes problemas têm tido cada vez mais atenção por parte dos investigadores, não se chegando, ainda, a resultados consensuais na comunidade científica. Este trabalho pretendeu contribuir para a procura de metodologias eficazes, no contexto dos problemas referidos.

Têm surgido na literatura, nomeadamente no contexto do *Data Mining*, alguns trabalhos realizados com o objectivo de produzir algoritmos que sejam robustos perante pontos isolados e quanto à detecção de classes de formas não esféricas. Como a eficácia de muitos destes algoritmos de classificação depende de alguns parâmetros de entrada, pretendeu-se que o método proposto nesta dissertação respondesse também a estes dois requisitos, não requerendo qualquer pressuposto sobre os dados.

A determinação do número de classes de um conjunto de dados numa análise não supervisionada, não é um problema de solução fácil devido à grande diversidade dos conjuntos de dados, nomeadamente quanto à forma das classes ou ao tipo de atributos. Neste contexto, foi apresentado o índice de *Lerman_Costa_Silva* que, associado a um método de partição baseado no algoritmo *k-médias* para atributos mistos (*k-prototype*), pretende identificar o número de classes existente no conjunto de dados. O método proposto mostrou ser tão eficiente como alguns índices existentes na literatura, no caso de atributos numéricos, mostrando-se superior quando aplicado a dados com

atributos mistos.

Usando alguns conceitos da teoria de grafos, foi definido um algoritmo de classificação não hierárquica baseado na coloração de grafos, com ajuste que reduz o número de cores, com o objectivo de que as classes encontradas, sejam classes homogéneas. O método de classificação foi associado a um índice de classificabilidade, também definido no contexto de grafos; em conjunto, pretendem identificar o número e a composição das classes do conjunto de dados. A eficácia do método foi constatada em conjuntos de dados artificiais, evidenciando a sua robustez no que diz respeito a conjuntos de dados com pontos isolados, com classes de formato não esférico e sem uma estrutura em classes.

Foram ainda apresentados alguns índices de avaliação de classes e partições, desenvolvidos no contexto da teoria de grafos e que, essencialmente, servem de suporte ao método proposto. Os índices de avaliação do isolamento absoluto e da densidade de classes permitem avaliar as classes obtidas pelo método, enquanto o índice de avaliação de partições permite decidir qual a partição a seleccionar de entre aquelas que apresentem o mesmo valor do índice de classificabilidade. Estes índices mostraram-se eficazes quando comparados com outros na literatura, mas sobretudo mostraram-se úteis na escolha de uma partição de entre as candidatas obtidas pelo método.

Os atributos foram ainda considerados de modo diferenciado, permitindo encontrar partições em sub-espaços de atributos. Esta metodologia mostrou-se eficiente quando aplicada a alguns conjuntos de dados reais e artificiais, quer quanto às partições φ-projectadas, quer quanto às partições consenso obtidas a partir das partições encontradas em sub-espaços de atributos.

Ao terminar esta dissertação ficou a sensação de que o trabalho realizado abre muitos caminhos para desenvolvimentos futuro.

No contexto dos atributos diferenciados, a primeira questão surge no modo de seleccionar os blocos de atributos. Uma das soluções poderá passar por determinar partições das variáveis (atributos), cujas classes fossem os blocos de atributos a serem usados para obter partições de indivíduos, pelo método com atributos diferenciados.

No caso das partições consenso obtidas a partir das partições em sub-espaços de atributos, podem estudar-se novas medidas de comparação de partições a serem optimizadas e novas heurísticas para a obtenção da partição consenso.

Poderia ainda ser explorado o caso de se considerarem grafos ponderados em que o custo de cada arco seria a dissemelhança entre os vértices correspondentes. Assim, a adjacência entre os vértices u e v deixaria de tomar valores em $\{0, 1\}$, consoante se verificasse $dist(u, v) < \alpha$ ou $dist(u, v) \geq \alpha$ respectivamente, para tomar o valor dado por $dist(u, v) - \alpha$, quando $dist(u, v) \geq \alpha$, que indicaria a "força" com que os vértices u e v seriam adjacentes.

O modelo de grafos aleatórios usado na definição das distribuições teóricas pressupõe independência entre arcos, o que nem sempre se verifica. Podem procurar-se alguns modelos, em alternativa aos clássicos, que tomem em consideração a possível dependência entre os arcos de um grafo aleatório, definindo-se sobre eles a distribuição dos índices apresentados neste trabalho.

Finalmente, o tempo de execução do método poderia ser significativamente reduzido implementando paralelismo, quer ao nível do algoritmo de coloração ([202], [83] e [85]), quer ao nível do ajuste.

Bibliografia

[1] C. Aaron. Normalisation et analyse des classes pour la classification sous hypothèse de connexité. In *11èmes Rencontres de la Société Francophone de Classification*, Bordéus, França, 2004.

[2] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 94–105, Seattle, Washington, 1998.

[3] N. Alon and M. Krivelevich. The concentration of the chromatic number of random graphs. *Combinatorica*, 17(3):303–313, 1997.

[4] K. Alsabti, S. Ranka, and V. Singh. An efficient k-means clustering algorithm. In *IPPS: 11th International Parallel Processing Symposium*, Orlando, Florida, 1998. IEEE Computer Society Press.

[5] D. Avis, C. Simone, and P. Nobili. On the chromatic polynomial of a graph. *Math. Program., Ser.B*, 92:439–452, 2002.

[6] H. Bacelar-Nicolau. *Analyse d'un algorithme de classification automatique.* Thèse de Doctorat de 3ème Cycle, ISUP, Université de Paris VI, 1972.

[7] H. Bacelar-Nicolau and F. Nicolau. Hierarchical clustering models and robustness. In *Actes des XXVI Journées de Statisque, A. S. U.*, pages 96–99, Neuchâtel, 1994.

[8] T. A. Bailey and R. Dubes. Cluster validity profiles. *Pattern Recognition*, 15(2):61–83, 1982.

[9] F. B. Baker. Stability of two hierarchical grouping techniques. Case I: Sensitivity to data errors. *Journal of the American Statistical Association*, 69(346):440–445, June 1974.

[10] F. B. Baker and L. J. Hubert. Measuring the power of hierarchical cluster analysis. *Journal of the American Statistical Association*, 70(349):31–38, March 1975.

[11] J. D. Banfield and A. E. Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, 49:803–821, 1992.

[12] J.-P. Barthélemy and M-F. Janowitz. A formal theory of consensus. *SIAM Journal on Discrete Mathematic*, 4:305–322, 1991.

[13] J.-P. Barthélemy and B. Leclerc. The median procedure for partitioning. In P. Hansen I.J. Cox and B. Julesz, editors, *Partitioning Data Sets, DIMACS Series in Discrete Mathematics and Theoretical Computer Sciences*, volume 19, pages 3–34. Springer, Berlin, 1995.

[14] J.-P. Barthélemy and B. Monjardet. The median procedure in cluster analysis and social choice theory. *Mathematical Social Science*, 1:235–267, 1981.

[15] M. L. Beale. Euclidean cluster analysis. *Bulletin of the International Statistical Institute*, 43(2):92–94, 1969.

[16] A. Ben-Dor, R. Shamir, and Z. Yakhini. Clustering gene expression patterns. *Journal of Comput. Biology*, 6(3-4):281–297, 1999.

[17] J. P. Benzécri. Construction d'une classification ascendante hiérarchique par la recherche en chaîne des voisins réciproques. *Cahiers de l'Analyse des Données*, 7(2):209–218, 1982.

[18] J. C. Bezdek and N. R. Pal. Some new indexes of cluster validity. *IEEE Transactions on Systems Management and Cybernetics- Part B: Cybernetics*, 28(3):301–315, June 1998.

[19] C. Biernacki. Précision sur les données et coude de la vraisemblance pour trouver le nombre de classes dans un mélange. *Revue de Statistique Appliquée*, 47(1):47–62, 1999.

[20] R. K. Blashfield. Mixture model tests of cluster analysis: Accuracy of four agglomerative hierarchical methods. *Psychological Bulletin*, 83(3):377–388, 1976.

[21] H. H. Bock. On some significance tests in cluster analysis. *Journal of Classification*, 2:77–108, 1985.

[22] H.-H. Bock and E.Diday. *Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data*. Springer, Heidelberg, 2000.

[23] B. Bollobás. Chromatic numer, girth and maximal degree. *Discrete Mathematics*, 24:311–314, 1978.

[24] B. Bollobás. The distribution of the maximum degree of a random graph. *Discrete Mathematics*, 32:201–203, 1980.

[25] B. Bollobás. Degree sequences of random graphs. *Discrete Mathematics*, 33:1–19, 1981.

[26] B. Bollobás. The diameter of random graphs. *Transactions of the American Mathematical Society*, 267(1):41–52, 1981.

[27] B. Bollobás. Vertices of given degree in a random graph. *Journal of Graph Theory*, 6:147–155, 1982.

[28] B. Bollobás. The chromatic number of random graphs. *Combinatorica*, 8(1):49–55, 1988.

[29] B. Bollobás. *Random Graphs, Second Edition*. Cambridge University Press, Cambridge, 2001.

[30] B. Bollobás and W. F. de La Vega. The diameter of random regular graphs. *Combinatorica*, 2(2):125–134, 1982.

[31] B. Bollobás and P. Erdös. Cliques in random graphs. *Math. Proc. Cambridge Philos. Soc.*, 80:419–427, 1976.

[32] B. Bollobás and V. Klee. Diameters of random bipartite graphs. *Combinatorica*, 4(1):7–19, 1984.

[33] J. N. Breckenridge. Replicating cluster analysis: Method, consistency, and validity. *Multivariate Bahavioral Research*, 24(2):147–161, April 1989.

[34] D. Brélaz. New methods to color the vertices of a graph. *Communications of the ACM*, 22:251–256, 1979.

[35] R. L. Brennan and R. J. Light. Measuring agreement when two observers classify people into categories not defined in advance. *British Journal of Mathematical and Statistical Psychology*, 27:154–163, 1974.

[36] P. Brito. *Métodos não hierárquicos de classificação - O método NHMEAN*. Tese de Mestrado, Faculdade de Ciências da Universidade de Lisboa, 1988.

[37] P. Brito, F. Sousa, and S. T. Pinto. Modèles VL en classification non-hiérarchique. In *Actas das 12èmes Rencontres de la Société Francophone de Classification*, Montréal, 2005.

[38] R. J. Brook and W. D. Stirling. Agreement between observers when the categories are not specified in advance. *British Journal of Mathematical and Statistical Psychology*, 37:271–282, 1984.

[39] R. B. Calinski and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3:1–27, 1974.

[40] D. Chakrabarti. AUTOPART: Parameter-free graph partitioning and outlier detection. In J.F. Boulicaut et al., editor, *Proceedings of PKDD*, volume LNAI 3202, pages 112–124. Springer-Verlag Berlin Heidelberg, 2004.

[41] G. Chartrand and L. Lesniak. *Graphs & Digraphs (Third Edition)*. Chapman & Hall/CRC, USA, 1996.

[42] A. Chaturvedi, K. Foods, P. Green, and J. D. Carroll. K-modes clustering. *Journal of Classification*, 18:35–55, 2001.

[43] P. Cheeseman and J. Stutz. Bayesian classification (AUTOCLASS): Theory and results. In *Advances in Knowledge Discovery and Data Mining*, pages 153–180. MIT Press, Cambridge, Mass., 1996.

[44] J. Cohen. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–9220, 1968.

[45] T. F. Coleman and J. J. Moré. Estimation of sparse jacobian matrices and graph coloring problems. *SIAM Journal of Numerical Analysis*, 20(1):187–209, 1983.

[46] J. C. Culberson. Iterated greedy graph coloring and the difficulty landscape. Technical Report TR92-07, Department of Computing Science, The University of Alberta, Edmonton, Alberta, Canada, 1992.

[47] B. V. Cutsem and B. Ycart. Indexed dendrograms on random dissimilarities. *Journal of Classification*, 15:93–127, 1998.

[48] J. Pinto da Costa and P. R. Rao. Central partition for a partition-distance and strong pattern graph. *REVSTAT - Statistical Journal*, 2(2):127–143, 2004.

[49] D. L. Davies and D. W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, April 1979.

[50] W. H. E. Day. Validity of clusters formed by graph-theoretic cluster methods. *Mathematical Biosciences*, 36:299–317, 1977.

[51] M. Delattre and P. Hansen. Bicriterion cluster analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-2(4):277–291, 1980.

[52] E. Diday. *Nouveaux Concepts et Nouvelles Méthodes en Classification Automatique*. Thèse d'état, University Paris VI, 1972.

[53] W. E. Donath and A. J. Hoffman. Lower bounds for the partitioning of graphs. *IBM Journal of Research and Development*, 17:420–425, 1973.

[54] R. O. Duda and P. E. Hart. *Pattern classification and scene analysis*. New York: Wiley, 1973.

[55] J. Dunn. A fuzzy relative of the isodata process and its use in detecting compact, well-separated clusters. *Journal of Cybernetics*, 3:32–57, 1974.

[56] M. Ester, H.-P. Kriegel, J. Sander, M. Wimmer, and X. Xu. Incremental clustering for mining in a data warehousing environment. In *Proceedings of 24th VLDB Conference*, pages 323–333, New York, USA, 1998.

[57] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In Portland, editor, *Proceedings of 2nd Int. Conference On Knowledge Discovery and Data Mining*, pages 226–231, 1996.

[58] V. Estivill-Castro and J. Yang. Fast and robust general purpose clustering algorithms. In *Pacific Rim International Conference on Artificial Intelligence*, pages 208–218, Melburne, Austrália, 2000.

[59] J. R. Allwright et al. A comparison of parallel graph coloring algorithms. Technical Report SCCS-666, Northeast Parallel Architecture Center, Syracuse University, 1995.

[60] D. Fasulo. An analysis of recent work on clustering algorithms. Technical report, University of Washington, 1999.

[61] U. Fayyad, C. Reina, and P. S. Bradley. Initialization of iterative refinement clustering algortithms. In *4th International Conference on Knowledge Discovery and Data Mining*, pages 194–198, New York, 1998.

[62] M. Fiedler. Algebraic connectivity of graphs. *Czechoslovak Math. Journal*, 23(98):298–305, 1973.

[63] I. Finocchi, A. Panconesi, and R. Silvestri. Experimental analysis of simple, distributed vertex coloring algorithms. In *Proceedings of the thirteenth annual ACM-SIAM Symposium on Discrete algorithms*, pages 606–615, San Francisco, California, 2002.

[64] L. Fisher and J. Van Ness. Admissible clustering procedures. *Biometrika*, 58:91–104, 1971.

[65] L. Fisher and J. W. Van Ness. Admissible discriminant. Programs in mathematical sciences, Technical Report 5, University of Texas, Dallas, 1971.

[66] W. D. Fisher. On grouping for maximum homogeneity. *J.A.S.A.*, 53:789–798, 1958.

[67] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steel, and P. Yanker. Query by image and video content: The QBIC system. *IEEE Computer*, 28:23–32, 1995.

[68] E. W. Forgy. Cluster analysis of multivariate data: efficiency vs. interpretability of classifications. In *WNAR Meeting, Univ. of California, Riverside, 1965; abstract in Biometrics, vol. 21, 3, page 768*, 1965.

[69] E. B. Fowlkes and C. L. Mallows. A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, 79(383):553–569, September 1983.

[70] C. Fraley and A. E. Raftery. How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal*, 41(8):578–588, 1998.

[71] O. Frank and D. Strauss. Markov graphs. *Journal of the American Statistical Association*, 81(395):832–842, 1986.

[72] A. Fred. Finding consistent clusters in data partitions. In Josef Kittler and Fabio Roli, editors, *Multiple Classifier Systems*, pages 309–318. Springer-Verlag, 2001.

[73] A. Fred. Clustering based on dissimilarity first derivatives. In *Proceedings of the 2nd Intl. Workshop on Pattern Recognition in Information Systems*, pages 257–266, 2002.

[74] A. Fred. Context-dependent clustering based on dissimilarity increments. In *Proceedings of the Portuguese Conference on Pattern Recognition*, Aveiro, Portugal, 2002.

[75] A. Fred and A. Jain. Data clustering using evidence accumulation. In *Proceedings of the 16th International Conference on Pattern Recognition*, pages 276–280, Quebec, Canada, 2002.

[76] A. Fred and A. Jain. Robust data clustering. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 128–133, Madison-Wisconsin, USA, 2003.

[77] J. Fridlyand and S. Dudoit. Applications of resampling methods to estimate the number of clusters and to improve the accuracy of a clustering method. Technical Report. No. 600, Stat. Berkeley, 2001.

[78] J. H. Friedman and L. C. Rafsky. Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *Annals of Statistics*, 7:697–717, 1979.

[79] A. M. Frieze. On the independence number of random graphs. *Discrete Mathematics*, 81:171–175, 1990.

[80] V. Ganti, J. Gehrke, and R. Ramakrishnan. CACTUS - clustering categorical data using summaries. In *Knowledge Discovery and Data Mining*, pages 73–83, 1999.

[81] J. A. Garcia, J. Fdez-Valdivia, F.J. Cortijo, and R. Molina. A dynamic approach for clustering data. *Signal Processing*, 44(2):181–196, 1994.

[82] M. R. Garey and D. S. Johnson. *Computers and Intractability.* H.W. Freeman, New York, 1979.

[83] A. H. Gebremedhin and F. Manne. Scalable parallel graph coloring algorithms. *Concurrency: Practice and Experience*, 12:1131–1146, 2000.

[84] D. Gibson, J. M. Kleinberg, and P. Raghavan. Clustering categorical data: An approach based on dynamical systems. *VLDB Journal: Very Large Data Bases*, 8(3–4):222–236, 2000.

[85] R. K. Gjertsen, M. T. Jones, and P. E. Plassmann. Parallel heuristics for improved, balanced graph colorings. *Journal of Parallel and Distributed Computing*, 37:171–186, 1996.

[86] E. Godehardt. *Graphs as Structural Models.* Friedr Vieweg & Sohn, Braunschweig / Wiesbaden, 1988.

[87] L. A. Goodman and W. H. Kruskal. Measures of association for cross classification. *Journal of the American Statistical Association*, 49:732–764, 1954.

[88] L. A. Goodman and W. H. Kruskal. Measures of association for cross classification IV: Simplification of asymptotic variances. *Journal of the American Statistical Association*, 67(338):415–421, June 1972.

[89] A. D. Gordon. Identifying genuine clusters in a classification. *Computational Statistics and Data Analysis*, 18:561–581, 1994.

[90] A. D. Gordon. Clustering validation. In C. Hayashi et al., editor, *Data Science, Classification, and Related Methods*, pages 22–39. Word Scientific Publ. River Edge, 1998.

[91] A. D. Gordon. *Classification - 2nd Edition.* Chapman & Hall/CRC, second edition, 1999.

[92] A. D. Gordon and M. Vichi. Partitions of partitions. *Journal of Classification*, 15:265–285, 1998.

[93] G. Grimmett and C. McDiarmid. On colouring random graphs. *Math. Proc. Cambridge Philos. Soc.*, 77:313–324, 1975.

[94] A. Guénoche. Enumération des partitions de diamètre minimum. *Discrete Mathematics*, 111:277–287, 1993.

[95] A. Guénoche. Partitions optimisées selon différents critères: évaluation et comparaison. *Mathématiques et Sciences Humaines*, 161:41–58, 2003.

[96] A. Guénoche. Clustering by density in any metric space. In M. Chavent, O. Dordan, C. Lacomblez, M. Langlais, and B. Patouille, editors, *Comptes rendus des 11émes Rencontres de la Société Francophone de Classification*, pages 200–203, Talence, France, September 2004.

[97] A. Guénoche. Clustering by vertex density in a graph. In Banks, House, LcMorris, Arabie, and Gaul, editors, *Classification, Clustering and Data Mining Applications*, pages 15–24. Springer, july 2004.

[98] A. Guénoche and H. Garreta. Representations and evaluation of partitions. In Jajuga, Sokolowski, and Bock, editors, *Classification, Clustering and Data Analysis. Recent Advances and Applications*, pages 131–138. Springer, Berlin Heidelberg, July 2002.

[99] S. Guha, R. Rastogi, and K. Shim. CURE: An efficient clustering algorithm for large databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 73–84, Seattle, Washington, USA, 1998.

[100] Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim. ROCK: A robust clustering algorithm for categorical attributes. *Information Systems*, 25(5):345–366, 2000.

[101] D. Gusfield. Partition-distance: a problem and class of perfect graphs arising in clustering. *Information Processing Letters*, 82(3):159–164, 2002.

[102] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2):107–145, 2001.

[103] M. Halkidi and M. Vazirgiannis. A data set oriented approach for clustering algorithm selection. In *Principles of Data Mining and Knowledge Discovery, 5th European Conference, PKDD*, volume 2168, Freiburg, Germany, 2001.

[104] M. Halkidi, M. Vazirgiannis, and Y. Batistakis. Quality scheme assessment in the clustering process. In *Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, pages 265–276, Lyon, France, 2000.

[105] P. Hansen and M. Delattre. Complete-link cluster analysis by graph coloring. *Journal of the American Statistical Association*, 17(362):397–403, June 1978.

[106] J. A. Hartigan. Statistical theory in clustering. *Journal of Classification*, 2:63–76, 1985.

[107] J. A. Hartigan and S. Mohanty. The RUNT test for multimodality. *Journal of Classification*, 9:63–70, 1992.

[108] R. L. Hoffman and A. K. Jain. Sparse decompositions for exploratory pattern analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(4):551–560, 1987.

[109] B. Hopkins. A new method for determining the type of distribution of plant-individuals. *Annals of Botany*, 18:213–226, 1954.

[110] Z. Huang. A fast clustering algorithm to cluster very large categorical data sets in data mining. In *SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*, pages 1–8, 1997.

[111] Z. Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining, Knowledge Discovery*, 2(3):283–304, 1998.

[112] L. Hubert. Inference procedures for the evaluation and comparison of proximity matrices. In J. Felsenstein, editor, *Numerical Taxonomy*, pages 209–228. Springer Verlag Berlin Heidelberg, 1983.

[113] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2:193–218, 1985.

[114] L. Hubert, P. Arabie, and J. Meulman. Modelling dissimilarity: Generalizing ultrametric and additive tree representations. *British Journal of Mathematical and Statistical Psychology*, 54:103–123, 2001.

[115] L. J. Hubert. Some applications of graph theory to clustering. *Psychometrika*, 39(3):283–309, September 1974.

[116] L. J. Hubert. Nominal scale response agreement as a generalized correlation. *British Journal of Mathematical and Statistical Psychology*, 30:98–103, 1977.

[117] L. J. Hubert and J. R. Levin. A general statistical framework for assessing categorical clustering in free recall. *Psychological Bulletin*, 83:1072–1080, 1976.

[118] M. Ichino and H. Yaguchi. Generalized Minkowski metrics for mixed feature-type data analysis. *IEEE Transactions on Systems, Man, and Cybernetics*, 24(4):698–708, 1994.

[119] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice-Hall, Englewood Cliffs, NJ, 1988.

[120] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[121] S. Janson, T. Luczak, and A. Rucinski. *Random Graphs*. Wiley Inter-Science, 2000.

[122] N. Jardine. Towards a general theory of clustering. *Biometrics*, 25:609–610, 1969.

[123] M. T. Jones and P. E. Plassmann. A parallel graph coloring heuristic. *SIAM Journal on Scientific Computing*, 14(3):654–669, 1993.

[124] R. Kannan, S. Vempala, and A. Vetta. On clusterings-good, bad and spectral. In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, 2000.

[125] T. Kanungo, D. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Wu. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):881–892, 2002.

[126] D. Karger, R. Motwani, and M. Sudan. Approximate graph coloring by semidefinite programming. *Journal of the ACM*, 45(2):246–265, 1998.

[127] M. Karoński and A. Ruciński. The origins of the theory of random graphs. In Rl L. Graham & J. Nesetrril, editor, *The Mathematics of Paul Erdos, I, Algorithms and Combinatorics*, volume 13, pages 311–336. Springer, Berlin, 1997.

[128] G. Karypis and V. Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on Scientific Computing*, 20(1):359–392, 1998.

[129] G. Karypis and V. Kumar. Multilevel algorithms for multi-constraint graph partitioning. In *Proceedings of SC98*, pages 1–13, 1998.

[130] G. Karypis and V. Kumar. Multilivel k-way hypergraph partitioning. Technical Report 98-036, Department of Computer Science & Engineering Army HPC Research Center, 1998.

[131] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: an Introduction to Cluster Analysis*. John Wiley & Sons, 1990.

[132] T. D. Klastorin. The p-median problem for cluster analysis: A comparative test using the mixture model approach. *Management Science*, 31(1):84–95, January 1985.

[133] B. Krieger and P. Green. A generalized Rand-index method for consensus clustering of separate partitions of the same data base. *Journal of Classification*, 16:63–89, 1999.

[134] W. J. Krzanowski and Y. T. Lai. A criterion for determining the number of groups in a data set using sum of squares clustering. *Biometrics*, 44:23–34, 1985.

[135] G. Lance and W. Williams. A general theory of classification sorting strategies: II - clustering systems. *The Computer Journal*, 10:271–277, 1969.

[136] G. N. Lance and W. T. Williams. A general theory of classificatory sorting strategies: II - clustering systems. *The Computer Journal*, 10(3), 1967.

[137] B. Leclerc and G. Cucumel. Consensus en classification: Une revue bibliographique. *Mathématiques et Sciences Humaines*, 100:109–128, 1987.

[138] I. Leenen, K. Leuven, I. Mechelen, and K. Leuven. An evaluation of two algorithms for hierarchical classes analysis. *Journal of Classification*, 18:57–80, 2001.

[139] I. C. Lerman. Sur l'analyse des données à une classification automatique. *Mathématiques et Sciences Humaines*, 32:5–15, 1970.

[140] I. C. Lerman. Sur la signification des classes issues d'une classification automatique de données. In J. Felsenstein, editor, *Vol G1 Numerical Taxaromy.* Springer-Verlag, 1983.

[141] I. C. Lerman, J. P. Costa, and H. B. Silva. Linéarisation d'un critère de classification en cas de données numériques et qualitatives nominales. In *Actas das VIIIèmes Rencontres de la Société Francophone de Classification*, Guadalupe, 2001.

[142] I. C. Lerman, J. P. Costa, and H. B. Silva. Validation of very large data sets clustering by means of a non parametric linear criterion. In Jajuga, Sokolowski, and Bock, editors, *Classification, Clustering and Data Analysis. Recent Advances and Applications*, pages 147–157. Springer, Berlin, Heidelberg, July 2002.

[143] I. C. Lerman, Ph. Peter, and H. Leredde. Principes et calculs de la méthode implantée dans le programme CHAVL (Classification hiérarchique par analyse de la vraisemblance des liens). *La Revue de Modulad*, 12:33–70, 1993.

[144] I. C. Lerman, Ph. Peter, and H. Leredde. Principes et calculs de la méthode implantée dans le programme CHAVL (Classification hiérarchique par analyse de la vraisemblance des liens)-deuxième partie. *La Revue de Modulad*, 13:63–90, 1994.

[145] F. R. Ling and G. G. Killough. Probability tables for cluster analysis based on a theory of random graphs. *Journal of the American Statistical Association*, 71:293–300, 1976.

[146] R. F. Ling. On the theory and construction of k-clusters. *Computer Journal*, 15:326–332, 1972.

[147] R. F. Ling. The expected number of componentes in random linear graphs. *The Annals of Probability*, 1(5):876–881, 1973.

[148] R. F. Ling. A probability theory of cluster analysis. *Journal of the American Statistical Association*, 68(341):159–164, March 1973.

[149] S.-Y. Lu. A tree-to-tree distance and its application to cluster analysis. *IEEE Trans. Pattern Anal. and Mach. Intelligence*, PAMI-1:219–224, 1979.

[150] M. Luby. A simple parallel algorithm for the maximal independent set problem. *SIAM, Journal of Computing*, 15(4):1036–1053, 1986.

[151] L. Lucera. The greedy coloring is a bad probabilistic algorithm. *Journal of Algorithms*, 12:674–684, 1991.

[152] T. Luczak. The chromatic number of random graphs. *Combinatorica*, 11(1):45–54, 1991.

[153] T. Luczak and J. C. Wierman. The chromatic number of random graphs at the double-jump threahold. *Combinatorica*, 9(1):39–49, 1989.

[154] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, 1967.

[155] O. L. Mangasarian. Mathematical programming in data mining. *Data Mining and Knowledge Discovery*, 1(2):183–201, 1997.

[156] H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18:50–60, 1947.

[157] D. Matula. On the complete subgraphs of a random graph. *Combinatory Math. and its Applications, Chapel Hill, N. C.*, pages 356–369, 1970.

[158] M. F. Mickevich. Taxonomic congruence: Rohlf and Sokal's misunderstanding. *Systematic Zool.*, 29:162–176, 1980.

[159] G. W. Milligan. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, 45:325–342, 1980.

[160] G. W. Milligan. An algorithm for generating artificial test clusters. *Psychometrika*, 50(1):123–127, 1981.

[161] G. W. Milligan. A Monte Carlo study of thirty internal criterion measures for cluster analysis. *Psychometrika*, 46(4):187–199, June 1985.

[162] G. W. Milligan and M. C. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179, 1985.

[163] G. W. Milligan and M. C. Cooper. A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioral Research*, 21:441–458, 1986.

[164] G. W. Milligan, S. C. Soon, and L. M. Sokol. The effect of cluster size, dimensionality, and the number of clusters on recovery of true cluster structure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-5(1):40–47, January 1983.

[165] B. Mirkin. Concept learning and feature selecting based on square-error clustering. *Machine Learning*, 35:25–39, 1999.

[166] N. Mishra, D. Ron, and R. Swaminathan. On finding large conjunctive clusters. In *Proceedings of the 16th COLT.*, pages 448–462, 2003.

[167] L. C. Morey and A. Agresti. The measurement of classification agreement: An adjustment to the Rand statistic for chance agreement. *Educational and Psychological Measurements*, 44:33–37, 1984.

[168] G. Bel Mufti. *Validation D'une Classe Par Estimation de sa Stabilité*. PhD. Thesis, Université Paris IX-Dauphine U.F.R. Mathématiques de la Décision, October 1998.

[169] D. W. Muller and G. Sawitzki. Excess mass estimates and tests for multimodality. *Journal of the American Statistical Association*, 86:738–746, 1991.

[170] S. Nascimento. *Fuzzy Clustering via proportional membership model*. IOS Press, Amsterdam, 2005.

[171] J. I. Naus and L. Rabinowitz. The expectation and variance of the number of componentes in random linear graphs. *The Annals of Probability*, 3(1):159–161, 1975.

[172] J. Ness. Recent results in clustering admissibility. In H. Bacelar-Nicolau, F. Costa Nicolau, and J. Janssen, editors, *Applied Stochastic Models and Data Analysis*, pages 19–29. Instituto Nacional de Estatistica, Lisboa, Portugal, 1999.

[173] J. W. Van Ness. Comment. *Journal of the American Statistical Association*, 78(383):569–576, September 1983.

[174] M. Newman, S. Strogatz, and D. Watts. Random graphs with arbitrary degree distribution and their applications. Technical Report 0007042, Santa Fe Institute, 2000.

[175] M. Newman, D. Watts, and S. Strogatz. Random graph models of social networks. In *Proceedings of the National Academy of Sciences*, volume 99, pages 2566 – 2572, 2002.

[176] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: analysis and an algorithm. In *Proceedings of NIPS*, pages 849–856, 2001.

[177] R. T. Ng and J. Han. Efficient and effective clustering methods for spatial data mining. In *Proceedings of the 20th VLDB Conference Santiago*, Santiago, Chile, 1994.

[178] F. Nicolau. *Critérios de análise classificatória hierárquica baseados na função distribuição*. Tese de doutoramento, Faculdade de Ciências, Universidade de Lisboa, 1980.

[179] F. Nicolau. *Métodos não-hierárquicos de classificação - O método das k-médias*. Prova complementar de doutoramento, Faculdade de Ciências, Universidade de Lisboa, 1981.

[180] F. Nicolau. Analysis of a non-hierarchic clustering method based on VL-similarity, Centro de Estatística e Aplicações / I.N.I.C., Nota n.18/85, 1985.

[181] F. Nicolau and P. Brito. Improvements in NHMEAN method. In E. Diday, editor, *Data Analysis, Learning Symbolic and Numerical Knowledge*, New York, 1989. Nova Science Publishers, Inc.

[182] N. R. Pal and J. C. Bezdek. On cluster validity for fuzzy c-means model. *IEEE Transactions on Fuzzy Systems*, 3(3):370–379, August 1995.

[183] N. R. Pal and J. Biswas. Cluster validation using graph theoretic concepts. *Pattern Recognition Society*, 30(6):847–857, 1997.

[184] Z. Palka. On the number of vertices of given degree in a random graph. *Journal of Graph Theory*, 18:167–170, 1984.

[185] D. Pelleg and A. Moore. X-means: Extending k-means with efficient estimation of the number of clusters. In *IPPS: 11th International Parallel Processing Symposium*. IEEE Computer Society Press, 1998.

[186] W. Polonik. Measuring mass concentrations and estimating density contour classes and excess mass approach. *Annals of Statistics*, 23:855–881, 1995.

[187] C. De Rahm. La CHA selon la méthode des voisins réciproques. *Cahiers de l'Analyse des Données*, 5(2):135–144, 1980.

[188] H. Ralambondrainy. A conceptual version of the k-means algorithm. *Pattern Recognition Letters*, 16:1147–1157, 1995.

[189] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, December 1971.

[190] S. Ray and R. Turi. Determination of number of clusters in k-means clustering and application in colour image segmentation. In *Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques*, pages 137–143, 1999.

[191] S. Régnier. Sur quelques aspects mathématiques des problèmes de la classification automatique. *I.C.C. Bulletin*, 4:175–191, 1965.

[192] B. Ripley. *Spatial Statistics*. Wiley, 1981.

[193] R. Robinson and A. Schwenk. The distribution of degrees in a large random tree. *Discrete Mathematics*, 12:359–372, 1975.

[194] F. J. Rohlf. Methods of comparing classifications. *Annual Review of Ecology and Systematics*, 5:101–113, 1974.

[195] F. J. Rohlf. Generalization of the gap test for the detection of multivariate outliers. *Biometrics*, 31:93–101, 1975.

[196] F. J. Rohlf. Consensus indices for comparing classifications. *Mathematical Biosciences*, 59:131–144, 1982.

[197] J. F. Rohlfand and R. D. Fisher. Tests for hierarchical structure in random data sets. *Systematic Zoology*, 17:407–412, 1968.

[198] V. Roth, M. Braun, T. Lange, and J. Buhmann. A resampling approach to cluster validation. In Bernd Ronz Wolfgang Hardle, editor, *Proceedings in Computational Statistics: 15th Symposium Held in Berlin (COMPSTAT2002)*, pages 123–128. Physica-Verlag, Heidelberg, Germany, 2002.

[199] P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[200] G. P. Rozál and J. A. Hartigan. The map test for multimodality. *Journal of Classification*, 11:5–36, 1994.

[201] E. Schikuta. Grid-clustering: An efficient hierarchical clustering method for very large data sets. Technical Report no. TR-96201, Institute of Applied Computer Science and Information Systems, University of Vienna, Austria, 1996.

[202] E. Shamir and E. Upfal. Sequential and distributed graph coloring algorithms with performance analysis in random graph spaces. *Journal of Algorithms*, 5:488–501, 1984.

[203] S. Sharma. *Applied Multivariate Techniques*. John Wiley & Sons, 1996.

[204] D. R. Shier. Testing for homogeneity using minimum spanning trees. *The UMAP Journal*, 3(3):273–283, 1982.

[205] H. B. Silva, J. P. Costa, and I. C. Lerman. Índice de validação de uma partição em análise classificatória. In *Actas da JOCLAD 2002: IX Jornadas de Classificação e Análise de Dados*, pages 44–47, Lisboa, 2002.

[206] S. P. Smith and A. K. Jain. Testing for uniformity in multidimensional data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 73–81, 1984.

[207] P. Smyth. Clustering using Monte Carlo cross-validation. In *Knowledge Discovery and Data Mining*, pages 126–133, 1996.

[208] R. R. Sokal and F. J. Rohlf. The comparison of dendrograms by objective methods. *Taxon*, 11:33–40, 1962.

[209] G. Soromenho. *Avaliação do Número de Componentes de uma Mistura. Aplicações em Classificação.* Tese de Doutoramento, Universidade Nova de Lisboa, 1993.

[210] F. Sousa. *Novas metodologias e validação em classificação hierárquica ascendente.* Tese de Doutoramento, Universidade Nova de Lisboa, Faculdade de Ciências e Tecnologia, 2000.

[211] D. Spielman and S. Teng. Spectral partitioning works: Planar graphs and finite elements meshes. In *Proceedings of the IEEE Symposium on Foundation of Computer Science*, 1996.

[212] D. J. Strauss. A model for clustering. *Biometrika*, 62:467–475, 1975.

[213] A. Strehl and J. Ghosh. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, 2002.

[214] Kendall & Stuart. *The Advanced Theory of Statistics*, volume 1. Third edition, 1976.

[215] M. J. Symons. Clustering criteria and multivariate normal mixtures. *Biometrics*, 37:35–43, 1981.

[216] S. Theodoridis and K. Koutroumbas. *Pattern Recognition.* Academic Press, 1999.

[217] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a dataset via the gap statistic. Technical Report 208, Department of Statistics, Stanford University, 2000.

[218] C. J. van Rijsbergen. A clustering algorithm. *Computer Journal*, 13:113–15, 1970.

[219] D. L. Wallace. Comment. *Journal of the American Statistical Association*, 78(383):569–576, September 1983.

[220] D. Welsh and M. Powell. An upper bound for the chromatic number of a graph and its application to timetabling problems. *Computer Journal*, 10:85–86, 1967.

[221] D. R. Wilson and T. R. Martinez. Improved heterogeneous distance functions. *Journal of Artificial Intelligence Research*, 6:1–34, 1997.

[222] M. P. Windham. Cluster validity for the fuzzy c-means clustering algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-4(4):357–363, July 1982.

[223] X. L. Xie and G. Beni. A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(8):841–847, August 1991.

[224] X. Xu, M. Ester, H.-P. Kriegel, and J. Sander. A distribution-based clustering algorithm for mining in large spatial databases. In *Proceedings of 14th International Conference On Data Engineering*, pages 209–218, 1998.

[225] G. Youness and G. Saporta. Une méthodologie pour la comparaison de partitions. *Rev. Statistique Appliquée*, LII(1):97–120, 2004.

[226] L. A. Zahn. Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Trans. Comput.*, 20:68–86, 1971.

[227] G. Zeng and R. C. Dubes. A test for spatial randomness based on k-NN distances. *Pattern Recognition Letters*, 3:85–91, 1985.

[228] T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: A new data clustering algorithm and its applications. *Data Mining and Knowledge Discovery*, 1(2):141–182, 1997.

[229] T. Zhang, R. Ramakrishnman, and M. Linvy. BIRCH: An efficient data clustering method for very large databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 103–114, 1996.

Anexo A

Interface do Programa

A implementação do método proposto nesta dissertação foi efectuada em *Microsoft Visual C++ 5.0* e a figura A.1 apresenta a *interface* do programa. Segue-se uma pequena descrição de cada um dos ítens:

- *Dados de Entrada.* Neste campo são pedidos o número de elementos a classificar (n), o número de atributos (t) e o nome do ficheiro que contém a matriz de dados $n \times t$ (<matriz_dados>.txt). O valor t representa o número total de atributos, tendo o ficheiro <matriz_dados>_Atrib.txt a caracterização dos mesmos. Assim, a linha i deste ficheiro contém a caracterização do atributo i, sendo {'n'} se for numérico e {'c' <tamanho>} se for categórico, onde <tamanho> indica o comprimento máximo das categorias do atributo i, para $1 \leq i \leq t$.

- *Parâmetro de Controlo.* Neste campo são pedidos os valores associados ao parâmetro de controlo, tal como descrito na Secção 4.4. Assim, se <Parâmetro de Controlo>= 1 então este parâmetro toma valores no intervalo $I = [d_{min}, d_{max}]$ onde d_{min} e d_{max} são respectivamente a dissemelhança mínima e máxima entre todos os pares de elementos. Se <Parâmetro de Controlo>= 0 então este parâmetro é dado pela média das dissemelhanças entre todos os pares de elementos. Se <Parâmetro de Controlo>= 3 então este parâmetro toma valores no intervalo $I = [\ell_{min}, \ell_{max}]$ onde ℓ_{min} e ℓ_{max} ($\ell_{min} \geq d_{min}$ e $\ell_{max} \leq d_{max}$) são dados pelo utilizador em <Limite Inferior> e <Limite Superior> respectivamente, <N Partes> indica o número de partes em que o intervalo I é dividido. Finalmente o utilizador pode atribuir ao <Parâmetro de Controlo> um outro valor, para o qual deseje obter a partição correspondente.

- *Opções de Execução.* Neste campo pode optar-se por uma execução ajustada do

método de partição baseado na coloração de grafos (MPCG), e pela estandardização das variáveis numéricas. Em alternativa à matriz de dados $n \times t$, pode ser fornecida a matriz de dissemelhanças $n \times n$ ou a matriz adjacente $n \times n$, sendo <Nome do Ficheiro> o nome do ficheiro que contém a respectiva matriz.

- *Método com Atributos não Diferenciados.* Neste campo é seleccionada a opção com atributos não diferenciados, selecciondo-se <Atributos Numéricos> se todos os t atributos forem numéricos ou <Atributos Mistos / Método das Dissemelhanças Huang> se os atributos forem de tipos mistos, sendo para este caso usado a dissemelhança de Huang (Fórmula (3.20), Secção 3.9). O índice de equilíbrio corresponde ao valor de γ da referida fórmula, e corresponde a 30% da média aritmética dos desvios padrão dos atributos numéricos se <Índice de Equilíbrio>=0, ou pode assumir outro valor escolhido pelo utilizador.

- *Método com Atributos Diferenciados.* Neste campo é seleccionada a opção com atributos diferenciados pela qual se escolhem partições φ-projectadas sendo o valor de φ dado por <Força de Ligação>, ou partições consenso, podendo ser medianas ou *medoids*.

- *Algoritmo de Coloração.* Neste campo é seleccionado o algoritmo de coloração de entre os descritos na Secção 4.7. O algoritmo escolhido em todos os exemplos apresentados ao longo da dissertação, foi o algoritmo de coloração sequencial com <Ordem Inicial>.

- *Avaliação de Classes e Partições.* Neste campo é(são) seleccionado o(s) índice(s) de avaliação da densidade ou isolamento de classes, ou o índice de avaliação de partições, descritos no Capítulo 5. É possível ainda obter os valores de alguns índices de validação encontrados na literatura, e que foram descritos também no capítulo 5. No caso do índice de densidade o <Param_ID>$\in \{0, 1, 2, 3\}$, correspondendo, respectivamente, aos seguintes valores (Secção 5.5.2): α, LC_i, $max\{\alpha, LC_i\}$ e $min\{\alpha, LC_i\}$. Para todas as classes para as quais foi calculado o índice de densidade considerou-se <Param_ID>= 0.

- *Método de Ajuste.* Neste campo pode optar-se, ou pelo ajuste de densidade (Secção 4.2.3) ou pelo ajuste hierárquico (Secção 5.8). Para este último caso se <Número de Classes Pretendido>= 0 então o ajuste é efectuado para k desconhecido.

- *Resultados.* Neste campo obtêm-se as três melhores partições encontradas pelo

método para as quais são apresentados os respectivos valores do índice de classificabilidade ($I\tilde{C}$), parâmetro de controlo (α) e número de classes.

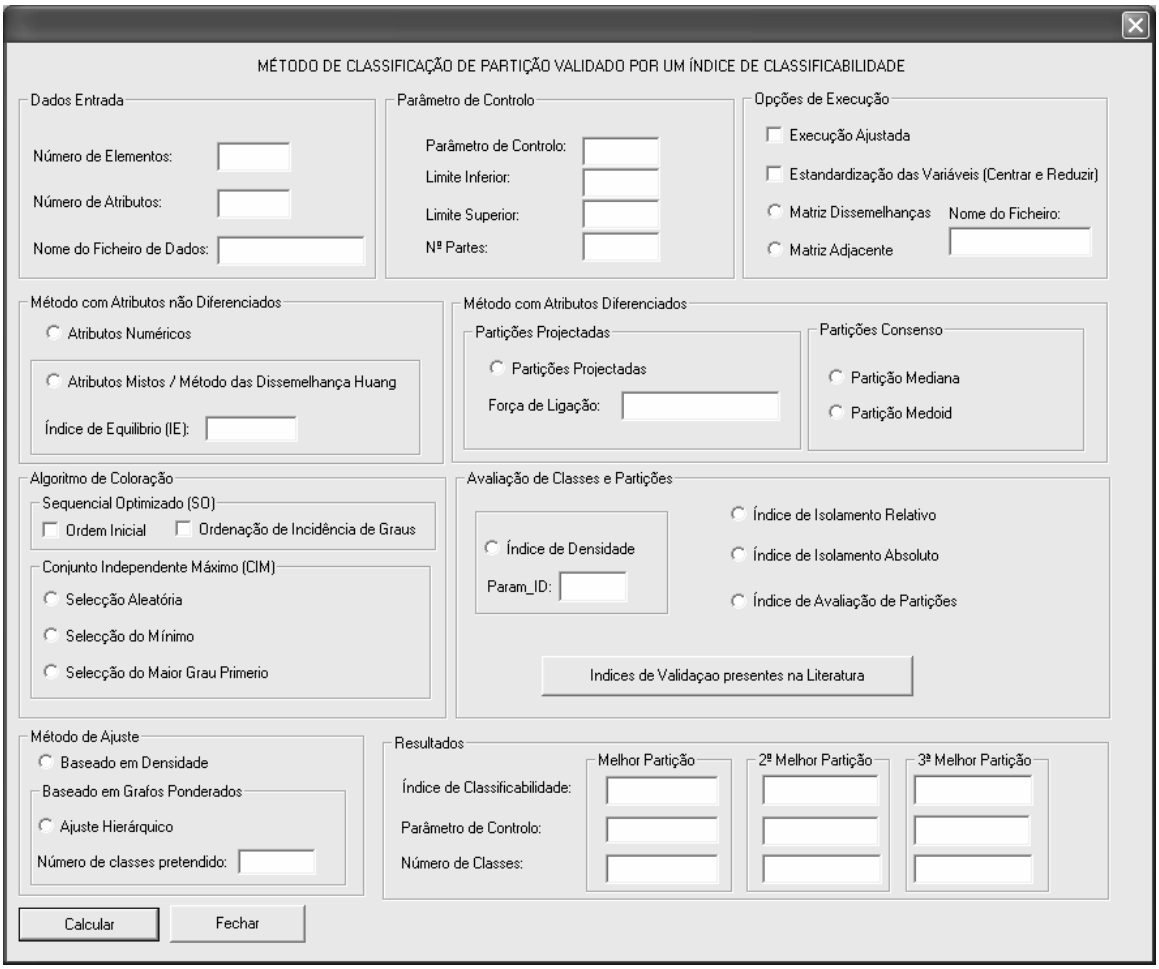

Figura A.1: *Interface* do programa que implementa o método proposto.