Agrupamento de Dados com Restrições

João Manuel Maia Duarte

Departamento de Engenharia Informática

Instituto Superior de Engenharia do Porto

Tese submetida para a obtenção do grau

Mestre

Novembro de 2008

Dedico esta dissertação à minha mãe e à minha irmã e peço a compreensão de ambas por não lhes ter dado a devida atenção e carinho no período em que decorreu o presente trabalho.

Agradecimentos

Gostaria de agradecer a todas as pessoas que de alguma forma contribuíram para a realização deste trabalho. Agradeço especialmente:

Ao meu orientador, Mestre Fernando Jorge Ferreira Duarte, por toda a ajuda prestada, pela disponibilidade que teve sempre e sobretudo pelo enorme trabalho de revisão desta dissertação;

À minha co-orientadora, Doutora Ana Luísa Nobre Fred, pelo vasto material bibliográfico disponibilizado e revisão desta dissertação;

Ao GECAD (Grupo de Investigação em Engenharia do Conhecimento e Apoio à Decisão) do Instituto Superior de Engenharia do Porto, por todos os meios disponibilizados para a realização deste trabalho;

Ao IT-Lisboa (Instituto de Telecomunicações - Lisboa) do Instituto Superior Técnico, pela bolsa de Iniciação à Investigação concedida no âmbito do projecto "*Ensemble Methods for Supervised and Semi-supervised Learning*", na qual desenvolvi muito do trabalho apresentado nesta dissertação;

À minha família por todo o seu apoio e compreensão ao longo deste último ano;

A todos os meus amigos, em especial ao grupo XT, que nunca se esqueceram de mim, apesar de muitas vezes ter negado a minha presença em eventos importantes.

A todos, muito obrigado!

Resumo

As técnicas de agrupamento de dados (classificação não supervisionada) são úteis em vários problemas de análise exploratória de dados, tomada de decisão, estruturação de documentos e segmentação de imagem, entre outros. O seu objectivo consiste na divisão de um conjunto de dados em vários grupos, em que dados semelhantes são colocados no mesmo grupo e dados dissemelhantes em grupos diferentes.

A combinação de agrupamentos de dados surgiu na última década com o intuito de melhorar a robustez e qualidade do agrupamento de dados, reutilizar soluções e agrupar dados de forma distribuída.

O agrupamento de dados com restrições tem como objectivo incorporar conhecimento *a priori* no processo de agrupamento de dados, com o intuito de aumentar a qualidade do agrupamento de dados e, simultaneamente, encontrar soluções apropriadas a tarefas ou interesses específicos.

Nesta dissertação, são estudados vários tipos de restrições usadas no agrupamento de dados, assim como os principais algoritmos de agrupamento de dados com restrições. São também desenvolvidas formas de combinar vários agrupamentos de dados usando restrições num agrupamento de dados final.

Com o propósito de comparar os algoritmos de agrupamento com restrições e de avaliar os métodos de combinação de agrupamentos de dados com restrições propostos, são realizados dois estudos comparativos usando conjuntos de dados de referência.

Palavras-Chave: *Aprendizagem Automática, Aprendizagem Semi-Supervisionada, Agrupamento de Dados, Agrupamento de Dados com Restrições.*

Abstract

Data clustering techniques (unsupervised classification) are useful in several problems of exploratory analysis, decision-making, documents structuring, image segmentation, among others. Its purpose is to partition a data set into several clusters, in which similar data is placed in the same cluster and dissimilar data in different clusters.

Cluster ensemble methods appeared in the last decade aiming to improve clustering robustness and quality, reuse clustering solutions and cluster data in a distributed way.

Constrained data clustering incorporates *a priori* knowledge in the clustering process, in order to improve data clustering quality and, simultaneously, find appropriated solutions to specific tasks or interests.

In this dissertation, several types of constraints related to data clustering are studied, as well as the main constrained data clustering algorithms.

We also developed new methods to combine several data clusterings using restrictions, into a final data clustering.

With the purpose of comparing the constrained data clustering algorithms and evaluating the proposed constrained cluster ensemble methods, two comparative studies are carried out using benchmark datasets.

Keywords: *Machine Learning, Semi-Supervised Learning, Data Clustering, Constrained Clustering.*

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Algoritmos

Notação

Conjuntos de Números

$\mathbb{N}$ - Conjunto dos números naturais, $\mathbb{N} = \{1, 2, ...\}$

$\mathbb{R}$ - Conjunto dos números reais

$x \in [a, b]$ - Intervalo $a \leq x \leq b$

$x \in]a, b]$ - Intervalo $a < x \leq b$

$x \in]a, b[$ - Intervalo $a < x < b$

$|C|$ - Cardinalidade de um conjunto C (em conjuntos finitos, o número de elementos)

Dados

$\mathcal{X}$ - Conjunto de dados

d - Dimensionalidade de $\mathcal{X}$

n - Número de objectos de dados de $\mathcal{X}$

K - Número de grupos do conjunto de dados

x_i - Objecto de dados $x_i \in \mathcal{X}$

l_i - Rótulo atribuído a x_i

P - Agrupamento/partição do conjunto de dados

C_k - k-ésimo grupo de um agrupamento de dados P

$\{\overline{x}_1, \cdots, \overline{x}_K\}$ - Centros dos K grupos que formam o agrupamento de dados P

$Rest_{=}$ - Conjunto de restrições de ligação obrigatória

$Rest_{\neq}$ - Conjunto de restrições de ligação proibida

$w_{=_{ij}}$ - Ponderação da restrição de ligação obrigatória entre x_i e x_j

$w_{\neq_{ij}}$ - Ponderação da restrição de ligação proibida entre x_i e x_j

$W_{=}$ - Conjunto de ponderações das restrições de ligação obrigatória

$W_{\neq}$ - Conjunto de ponderações das restrições de ligação proibida

P^i - i-ésimo agrupamento de um conjunto de agrupamentos

K^i - Número de grupos existentes no agrupamento de dados P^i

N - Número de agrupamentos de dados a combinar

$\mathcal{P}$ - Conjunto de agrupamentos de dados, $\mathcal{P} = \{P^1, \cdots, P^N\}$

$\mathcal{C}$ - Conjunto das K classes a que um objecto pode pertencer, $\mathcal{C} = \{c_1, \cdots, c_K\}$

$V(x_i)$ - Conjunto de objectos vizinhos do objecto x_i

V_p - p-ésimo conjunto de vizinhança

Vectores, Matrizes e Normas

A^{T} - Matriz transposta da matriz A

$det(A)$ - Determinante da matriz A

$\langle x_i, x_j \rangle$ - Produto interno entre x_i e x_j

$||\cdot||$ - Norma euclidiana, $||x|| = \sqrt{\langle x, x \rangle}$

$||\cdot||_p$ - Norma no espaço l^p, $||x||_p = (\sum_{i=1}^{N} |x_i|^p)^{\frac{1}{p}}$

co_assoc - Matriz de co-associações entre os n objectos do conjunto de dados

Funções

$d(x_i, x_j)$ - Distância entre os objectos x_i e x_j

g_i - Índice do grupo mais próximo de x_i

h_i - Índice do grupo mais próximo do centro de grupo $\overline{x}_{l_i}$ a que pertence x_i

$I(\cdot)$ - Função que devolve um caso a expressão seja verdadeira, devolvendo 0 no caso contrário

$\jmath obj$ - Função-objectivo obj

Probabilidades

$Pr(\cdot)$ - Probabilidade do evento $(\cdot)$

$p(\cdot)$ - Densidade de probabilidade de $(\cdot)$

Capítulo 1

Introdução

1.1 Enquadramento

A aprendizagem automática consiste no estudo de algoritmos que melhorem automaticamente os seus desempenhos através da experiência. Tradicionalmente, existem duas grandes áreas na aprendizagem automática: a aprendizagem supervisionada e a aprendizagem não supervisionada.

A aprendizagem supervisionada tem como objectivo aprender o mapeamento entre os atributos de um objecto de dados x e respectivo rótulo l, usando para isso um conjunto de treino constituído por pares (x_i, l_i), em que l_i é o atributo-alvo (classe) a que pertence o objecto x_i. O mapeamento $x \rightarrow l$ permite classificar novos objectos na classe respectiva, tendo em consideração a experiência resultante do conjunto de treino.

Na aprendizagem não supervisionada não são conhecidos os rótulos dos n objectos que formam o conjunto de dados $\mathcal{X} = \{x_i, \cdots, x_n\}$. O seu objectivo consiste em encontrar estruturas interessantes no conjunto de dados $\mathcal{X}$. Uma das áreas mais importantes na aprendizagem não supervisionada é o agrupamento de dados. O agrupamento de dados tem como objectivo dividir os objectos x_i de um conjunto de dados $\mathcal{X}$ em K grupos, tal que, objectos pertencentes ao mesmo grupo possuam características semelhantes e objectos agrupados em grupos diferentes tenham características distintas.

Entre a aprendizagem supervisionada e a aprendizagem não supervisionada encontra-se a aprendizagem semi-supervisionada. Para além dos valores para os atributos dos objectos $x_i \in \mathcal{X}$, é fornecida alguma informação de supervisão, não tendo esta informação de ser obrigatoriamente dada para todos os objectos x_i do conjunto de dados $\mathcal{X}$. Existem duas perspectivas para a aprendizagem semi-supervisionada: a classificação/regressão de dados semi-supervisionada e o agrupamento de dados semi-supervisionado, também denominado agrupamento de dados com restrições. A primeira perspectiva assume que o conjunto de dados $\mathcal{X} = \{x_i, \cdots, x_n\}$ pode ser divido em dois conjuntos $\mathcal{X}_r = \{x_i, \cdots, x_{num\_rot}\}$ e $\mathcal{X}_{nr} = \{x_i, \cdots, x_{num\_n\_rot}\}$, em que $\mathcal{X}_r$ corresponde ao conjunto de num_rot objectos para os quais existem rótulos $\{l_1, \cdots, l_{num\_rot}\}$ e $\mathcal{X}_{nr}$ corresponde ao conjunto de objectos para os quais não é conhecida a classe. O objectivo

consiste em fazer o mapeamento $x \rightarrow l$ para classificar novos objectos de dados, usando ambos os conjuntos de objectos $\mathcal{X}_r$ e $\mathcal{X}_{nr}$. A segunda perspectiva usa, para além do conjunto de objectos $\mathcal{X}$, informação *a priori* sobre a estrutura dos dados. Esta informação é representada na forma de restrições e é usada para influenciar a descoberta da estrutura dos dados. Geralmente, o uso de restrições permite que a estrutura de dados encontrada vá de encontro a interesses específicos, contrariamente às soluções encontradas pelos algoritmos de aprendizagem não supervisionada, em que são optimizados critérios gerais a todos os problemas. Como principal objectivo desta dissertação, pretende-se estudar várias formas de incorporar restrições na aprendizagem semi-supervisionada, mais precisamente, no agrupamento de dados com restrições.

Nos últimos anos, têm sido estudadas várias formas de combinar soluções, tanto da aprendizagem supervisionada como da não supervisionada, com o intuito de melhorar o desempenho das tarefas de classificação e agrupamento de dados. Nestas abordagens, são geradas várias soluções usando um ou vários algoritmos de aprendizagem, sendo as soluções combinadas em apenas numa solução de consenso. Espera-se que a solução resultante da combinação de várias soluções herde as boas características das soluções que a originaram, resultando numa solução final de qualidade superior. Nesta dissertação, são desenvolvidos métodos que incorporam restrições para combinar vários agrupamentos de dados.

1.2 Objectivos e Principais Contribuições

Esta dissertação tem como objectivo principal o estudo de técnicas que permitam a incorporação de conhecimento *a priori* no processo de agrupamento de dados, com o intuito de que o agrupamento de um conjunto de dados vá de encontro a tarefas ou interesses particulares. Assim, espera-se que a solução obtida por um algoritmo de agrupamento de dados, que use conhecimento de domínio relativo ao conjunto de dados a agrupar, seja mais relevante e vantajosa em aplicações específicas. As principais contribuições deste trabalho são:

Revisão do estado da arte em Agrupamento de Dados com Restrições. Estudo dos vários tipos de restrições usadas no agrupamento de dados, assim como dos principais algoritmos de agrupamento de dados com restrições. São apresentados os conceitos fundamentais da aprendizagem supervisionada, não supervisionada e semi-supervisionada para contextualizar o agrupamento de dados com restrições.

Quatro propostas para a combinação de agrupamentos de dados usando restrições. São propostas duas versões modificadas do método de Acumulação de Evidências [35], um método que transforma o conjunto de agrupamento de dados num novo conjunto de dados, aplicado em seguida o algoritmo de agrupamento de dados categóricos COP-COBWEB [85], e um método baseado na optimização da Medida de Consistência de Grupos [27] com penalização de violações de restrições usando um algoritmo genético.

Estudo comparativo entre vários algoritmos de agrupamento de dados com restrições. É realizado um estudo comparativo entre vários dos algoritmos de agrupamento de dados com restrições apresentados neste dissertação, com o intuito de avaliar o desempenho destes algoritmos em variados conjuntos de dados.

Avaliação dos métodos de combinação de agrupamentos propostos. É também realizado um estudo comparativo entre os métodos de combinação de agrupamentos de dados com restrições propostos, com o objectivo de validar o desempenho dos mesmos.

1.3 Guia de Leitura

Nesta secção é apresentado o guia de leitura do presente documento. Esta dissertação é composta por 7 capítulos, sendo cada capítulo descrito resumidamente de seguida:

- Neste primeiro capítulo, *Introdução*, é efectuado o enquadramento do tema desta dissertação, o Agrupamento de Dados com Restrições, sendo também apontados os principais objectivos deste trabalho, bem como, as principais contribuições.

- No segundo capítulo, *Aprendizagem Automática*, pretende-se introduzir alguns conceitos fundamentais da Aprendizagem Automática. Neste capítulo são apresentadas sucintamente as áreas: Aprendizagem Supervisionada, focando o tema Classificação de Dados; a Aprendizagem Não Supervisionada, mais precisamente o Agrupamento de Dados; e, finalmente, a Aprendizagem Semi-Supervisionada, sendo efectuada a distinção entre a Classificação de Dados Semi-Supervisionada e o Agrupamento de Dados com Restrições.

- No terceiro capítulo, *Agrupamento de Dados com Restrições*, é apresentada uma visão geral dos vários tipos de restrições que podem ser incorporadas no agrupamento de dados. Neste capítulo são também descritos dois métodos para a aquisição de restrições.

- No quarto capítulo, *Algoritmos de Agrupamento de Dados com Restrições*, são descritos os principais algoritmos de agrupamento de dados que possuem a capacidade de incluír restrições no processo de agrupamento, estando os algoritmos de agrupamento de dados categorizados considerando as ideias e intuições em que se baseiam.

- No quinto capítulo, *Combinação de Agrupamentos de Dados com Restrições*, é introduzido o tema da combinação de soluções resultantes de algoritmos da Aprendizagem Automática. Inicialmente, é discutido o tema da combinação de classificadores de dados, sendo descritos os problemas fundamentais da aplicação individual de classificadores de dados e as principais abordagens para a combinação de vários classificadores. Em seguida, é introduzido o tema da combinação de agrupamentos de dados, sendo apresentadas as principais vantagens da sua aplicação e as abordagens de combinação de agrupamentos de dados mais importantes. Finalmente, são apresentadas quatro abordagens para a combinação de agrupamentos de dados com a capacidade de incorporarem restrições.

- No sexto capítulo, *Avaliação de Algoritmos de Agrupamento de Dados e Métodos de Combinação*, é realizado um estudo comparativo entre alguns dos algoritmos de agrupamento de dados com restrições, descritos no quarto capítulo desta dissertação, sendo usado o bem conhecido algoritmo K-médias como referência. É também efectuado outro estudo comparativo, com o intuito de se avaliar o desempenho dos métodos de combinação propostos, tendo como referência o método de combinação de agrupamentos de dados Acumulação de Evidências.

- As conclusões deste trabalho são apresentadas no sétimo e último capítulo, *Conclusões*. Neste capítulo são descritas as principais limitações deste trabalho, sendo apresentadas direcções para trabalho futuro relacionado com o agrupamento de dados com restrições.

Capítulo 2

Aprendizagem Automática

2.1 Introdução

A aprendizagem automática é uma área de investigação que tem o objectivo de dotar o computador com a capacidade de aprendizagem, estudando para isso algoritmos e técnicas que permitam ao computador aperfeiçoar-se no desempenho de uma determinada tarefa, sem que seja necessária intervenção humana. Duas das grandes áreas de estudo da aprendizagem automática são a aprendizagem supervisionada e a aprendizagem não supervisionada. A aprendizagem supervisionada, mais particularmente a classificação de dados, tem o âmbito de proporcionar previsões baseadas na análise de conjuntos de dados. Na classificação de dados, cada objecto do conjunto de dados tem associado um valor para um atributo alvo, isto é, a classe a que o objecto pertence. As suas técnicas têm várias aplicações, tais como reconhecimento de caracteres, reconhecimento de imagens, filtragem de *spam*, diagnóstico médico, previsão de riscos financeiros, entre muitas outras. Na aprendizagem não supervisionada não se sabe qual a classe de cada objecto, o que torna a a aprendizagem bem mais complicada. No entanto, as técnicas não supervisionadas baseiam-se na noção de que o objectivo da máquina consiste em elaborar representações dos dados de entrada que possam ser úteis na tomada de decisão e na previsão de novos dados.

2.2 Aprendizagem Supervisionada

A aprendizagem supervisionada é uma área da aprendizagem automática que tem como objectivo criar uma função ou regra de decisão, a partir de um conjunto de dados de treino, que permita efectuar uma previsão/decisão sobre novos dados. Cada objecto desse conjunto de dados tem associado um atributo alvo, a classe do objecto. A tarefa de um algoritmo de classificação é prever a classe para um novo objecto de dados. Para isso, o algoritmo de aprendizagem usa um critério que permite a generalização a novos dados, baseada nas observações do conjunto de treino.

Um classificador é uma função f que atribui a um objecto x_i de um conjunto de dados $\mathcal{X}$ ($x_i \in \mathcal{X}$) caracterizado por d atributos, uma classe $c_k \in \mathcal{C}$:

$$f : \mathbb{R}^d \rightarrow \mathcal{C} \tag{2.1}$$

em que $\mathcal{C} = \{c_1, c_2, \cdots, c_K\}$ representa o conjunto de todas as K classes a que um objecto de dados pode pertencer.

Existem vários algoritmos para treinar classificadores de dados, isto é, aprender a função ou regra de decisão. Algumas das principais abordagens para a classificação de dados são descritas no texto que se segue.

Métodos Estatísticos. Muitas dos algoritmos clássicos de classificação de dados baseiam-se em modelos estatísticos, tais como, a análise discriminante linear e quadrática de dados, em que se pressupõe que os objectos de dados de cada classe são gerados segundo uma distribuição Gaussiana multivariada. Um método bastante conhecido é o algoritmo de classificação K-vizinhos-mais-próximos, que calcula a distância do objecto que se pretende classificar, x_i, a todos os objectos do conjunto de dados, classificando x_i com a classe mais representada nos K objectos mais próximos de x_i. Este método pode ser considerado estatístico uma vez que estima localmente densidades de probabilidades.

Árvores de Decisão. Existem algoritmos de classificação que se baseiam na definição de regras, seguindo uma estrutura em árvore, que definem qual a classe de um objecto de dados. Dois exemplos destes algoritmos de classificação são o CART (*Classification and Regression Tree*) [16] e o ID3 [73] (*Iterative Dichotomiser 3*).

Redes Neuronais Artificiais. As redes neuronais artificiais baseiam-se no funcionamento do cérebro humano e surgiram com a ideia de modelar matematicamente as capacidade intelectuais do ser humano. A estrutura de redes neuronais mais utilizada para a classificação de dados é a rede perceptrão multi-camada, ou MLP (*MultiLayer Perceptron*), existindo diversos algoritmos para treinar a rede, tal como o algoritmo de retropropagação [74].

Máquinas de Suporte Vectorial. As Máquinas de Suporte Vectorial (*Support Vector Machines* - SVM) têm como objectivo encontrar um plano de decisão que separe os objectos de dados com classes diferentes. Desta forma, para se classificar um novo objecto de dados basta saber a sua posição relativamente ao plano de decisão [21].

Combinação de Classificadores. Outra abordagem para a classificação de dados consiste na combinação de vários classificadores. O principal objectivo consiste em aumentar o desempenho da classificação de dados, usando para isso diversos classificadores obtidos por um ou vários algoritmos de classificação. Este assunto será abordado com mais detalhe na secção 5.2 do capítulo 5 desta dissertação.

2.3 Aprendizagem Não Supervisionada

Na aprendizagem não supervisionada, pretende-se que o computador aprenda sem que para isso lhe sejam fornecidos exemplos rotulados, contrariamente ao que acontece na aprendizagem supervisionada. Uma das técnicas mais usadas da aprendizagem não supervisionada é o agrupamento de dados, também denominada classificação não supervisionada. No resto desta dissertação, classificação de dados refere-se a classificação supervisionada de Dados enquanto que a classificação não supervisionada será sempre referida por agrupamento de dados.

As técnicas de agrupamento de dados são bastante úteis para descobrir distribuições com significado ou classes em dados. O problema do agrupamento de dados consiste na divisão de um conjunto de dados em grupos, de forma a colocar objectos de dados semelhantes ou "próximos"no mesmo grupo e objectos de dados dissemelhantes ou "afastados" em grupos diferentes. A sua aplicação é variada: genética, biologia, engenharia, economia, entre outros. Devido à não existência de classes predefinidas nem qualquer outro tipo de informação que indique a estrutura da informação a analisar, o processo de agrupamento de dados pode resultar em agrupamentos de dados diferentes, dependendo do critério especificado para o processo. Existe ainda a necessidade de pré-processar os dados a examinar no sentido de se seleccionar apenas a informação efectivamente relevante para o problema em causa. As fases principais do agrupamento de dados são:

- Selecção dos atributos. O objectivo é seleccionar os atributos do conjunto de dados realmente importantes para a análise que se pretende efectuar.

- Algoritmo de agrupamento de dados. Nesta fase escolhe-se um algoritmo de agrupamento de dados para encontrar a estrutura do conjunto de dados. Um algoritmo de agrupamento de dados é caracterizado por uma medida de proximidade e por um critério de agrupamento de dados que induzem o resultado final (agrupamento de dados).

- A medida de proximidade quantifica a semelhança entre dois objectos de um conjunto de dados. Normalmente, todos os atributos contribuem de igual forma para a definição da proximidade entre os objectos.

- O critério de agrupamento de dados pode ser visto como uma função de custo ou outro tipo de regra e determina a forma como o agrupamento do conjunto de dados é efectuado. Para a escolha do critério de agrupamento de dados deve-se ter em atenção, se possível, a forma dos grupos.

- Validação dos resultados. A qualidade dos resultados obtidos por um algoritmo de agrupamento de dados é avaliada através de técnicas e critérios apropriados, denominados índices de validação de agrupamentos de dados. Como os algoritmos de agrupamento de dados não têm conhecimento prévio dos grupos presentes no conjunto de dados, é necessário avaliar a correcção do agrupamento de dados final.

- Interpretação de resultados. Geralmente, os resultados do processo de agrupamento de dados são integrados com evidências resultantes de outras experiências ou análises para que possam ser tiradas conclusões.

As principais aplicações do agrupamento de dados são:

Compressão de dados. Em muitos casos, o volume de informação é muito grande e o processamento da mesma torna-se computacionalmente muito exigente. O agrupamento de dados pode ser utilizado para dividir o conjunto de dados num número apropriado de grupos. Em vez de se processar o conjunto de dados na totalidade, poder-se-á processar apenas dados representativos dos grupos obtidos, comprimindo-se desta forma o conjunto de dados inicial.

Geração de hipóteses. O agrupamento de dados pode ser utilizado de forma a possibilitar a inferência de algumas hipóteses respeitantes ao conjunto de dados.

Teste de hipóteses. Neste caso o agrupamento de dados é utilizado com o intuito de verificar a validade de uma determinada hipótese.

Previsão baseada em grupos. Os grupos resultantes do processo de agrupamento sobre um conjunto de dados são caracterizados pelos atributos dos objectos que os constituem. Surgindo um novo objecto, pode ser classificado com base na sua similaridade com os grupos obtidos anteriormente.

Biologia. Na biologia, o agrupamento de dados é útil na categorização de genes com comportamentos semelhantes e pode ser utilizado na verificação de estruturas inerentes a determinadas populações.

Análise de informação espacial. Devido à grande quantidade de informação que pode ser obtida através de imagens de satélite e sistemas de informação geográfica (GIS), entre outros, é caro e difícil examinar toda esta informação em detalhe. O agrupamento de dados pode ser utilizado para identificar e entender padrões interessantes nesses conjuntos de dados.

Web mining. Neste caso, o agrupamento de dados é utilizado na descoberta de grupos de documentos relacionados e alojados na *Web*, permitindo assim o acesso a esses documentos de uma forma eficiente.

De seguida, indicam-se alguns dos principais algoritmos de agrupamento de dados, apresentados segundo a taxonomia definida por Duarte [27]. As abordagens de agrupamento de dados podem ser divididas em cinco categorias: abordagens de partição, hierárquicas, baseadas em densidade, baseadas em grelha e baseadas em modelos.

Abordagens de partição. Os algoritmos de agrupamento de dados de partição procuram estruturar um conjunto de dados $\mathcal{X}$ num agrupamento de dados $P = \{C_1, \cdots, C_K\}$ com K grupos, optimizando uma função-objectivo, $f : P \rightarrow \mathbb{R}$, que procura reunir objectos semelhantes no mesmo grupo e colocar objectos dissemelhantes em grupos diferentes. Inicialmente, os algoritmos de agrupamento desta categoria criam um primeiro agrupamento de dados e, em seguida, efectuam iterativamente a realocação de objectos de um grupo para outro, com o intuito de minimizar a função-objectivo f. O erro quadrático é a função-objectivo mais utilizada para o efeito

$$f = \sum_{k=1}^{K} \sum_{x_i \in C_k} ||x_i - \overline{x}_k||^2 \tag{2.2}$$

em que $||x_i - \overline{x}_k||$ consiste na distância euclidiana entre o objecto x_i e o centro de grupo mais próximo $\overline{x}_k$.

O algoritmo de agrupamento de dados K-médias [58] é o algoritmo mais conhecido e usado das abordagens de partição e, provavelmente, de todas as abordagens. Este algoritmo recebe como parâmetro de entrada o número de grupos pretendido, K, e inicialmente escolhe de forma aleatória K objectos do conjunto de dados $\mathcal{X}$ como os centros (centróides) iniciais de cada grupo, $\{\overline{x}_1, \cdots, \overline{x}_K\}$. Após este primeiro passo, o algoritmo K-médias resume-se a atribuir cada objecto x_i ao grupo C_k cujo centro $\overline{x}_k$ se encontra mais próximo e em seguida actualizar cada centro de grupo, $\overline{x}_k$, como sendo o vector médio dos $|C_k|$ objectos associados a esse grupo, $\overline{x}_k = \frac{\sum_{x_i \in C_k} x_i}{|C_k|}$. Este processo repete-se até que não exista qualquer modificação nos grupos de uma iteração para a seguinte, garantindo que a função-objectivo (a apresentada na equação 2.2) converge para um óptimo (local).

Os algoritmos de agrupamento K-medóides são outros algoritmos desta categoria. Enquanto que no algoritmo K-médias, cada grupo é representado pelo vector médio dos objectos que contém, nos algoritmos K-medóides são apenas considerados vectores relativos a objectos para representar cada grupo, denominados *medóides*, correspondendo esses vectores aos objectos mais centralmente localizados em cada grupo. Inicialmente, são escolhidos aleatoriamente K objectos para serem os medóides dos K grupos. Cada um dos restantes objectos de dados são associados ao grupo cujo medóide é mais próximo, formando o primeiro agrupamento de dados. Em seguida, trocam-se iterativamente os medóides de cada grupo com o objectivo de optimizar a função-objectivo (frequentemente o erro quadrático apresentado na equação 2.2) até que seja encontrado um critério de paragem. Os algoritmos de agrupamento PAM (*Partitioning Around Medoids*)[49] e CLARANS (*Clustering LArge Application based upon RANdomized Search*)[65] são exemplos de algoritmos baseados em K-medóides.

Abordagens hierárquicas. Os algoritmos de agrupamento hierárquicos estruturam os objectos de um conjunto de dados numa hierarquia, representada na forma de árvore, denominada dendrograma.

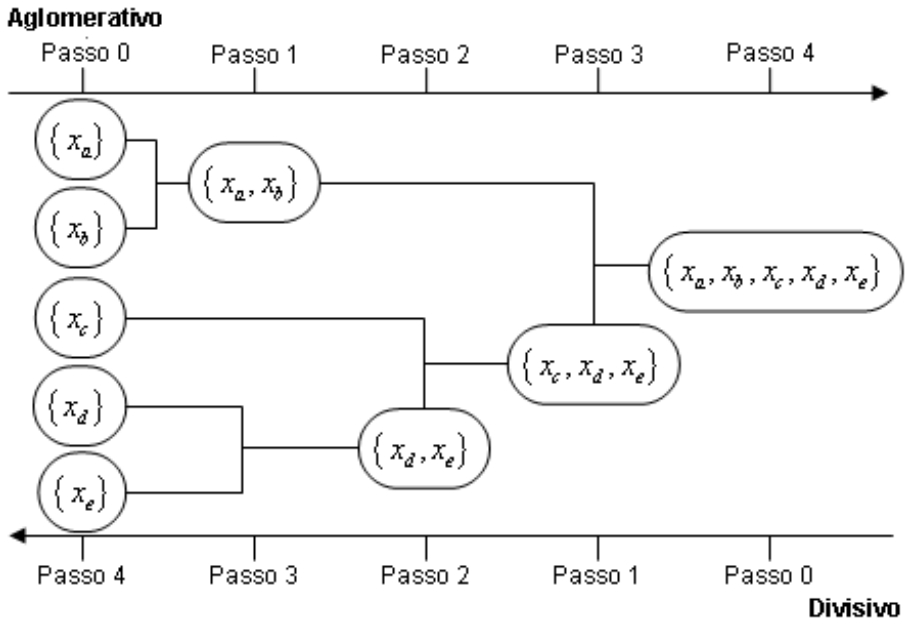

Figura 2.1: Agrupamento de dados hierárquico - Da esquerda para a direita é realizado agrupamento aglomerativo enquanto que da direita para a esquerda o agrupamento é divisivo. (Figura retirada de [27]).

Classicamente, os algoritmos de agrupamento hierárquicos podem ser classificados como aglomerativos ou divisivos. Nos algoritmos aglomerativos, cada um dos n objectos $x_i \in \mathcal{X}$ forma inicialmente um grupo $C_i = \{x_i\}$. De seguida, são fundidos sucessivamente os dois grupos mais similares (tendo em conta uma determinada função-objectivo) até que todos os objectos pertençam ao mesmo grupo. O funcionamento dos algoritmos hierárquicos aglomerativos encontra-se ilustrado na figura 2.1, seguindo os passos da esquerda para a direita. Nos algoritmos divisivos acontece o inverso. Inicialmente, todos os objectos encontram-se no mesmo grupo, sendo os grupos divididos sucessivamente em dois, até que cada grupo seja composto por apenas um objecto de dados. A figura 2.1 ilustra também o funcionamento dos algoritmos de agrupamento hierárquicos divisivos, seguindo os passos da direita para a esquerda. Os algoritmos aglomerativos são mais usados que os divisivos, já que são computacionalmente menos dispendiosos. Os algoritmos ligação simples [76], ligação completa [51], média de grupo [76], ligação centróide [76] e ligação de Ward [89] são exemplos clássicos da abordagem hierárquica aglomerativa.

O algoritmo CURE (*Clustering Using REpresentatives*) [40] é também um algoritmo de agrupamento hierárquico aglomerativo, mas segue uma estratégia diferente. Contrariamente aos algoritmos anteriormente mencionados, que usam apenas um vector representativo por cada grupo (objecto de dados ou centróide) para calcular as similaridades entre grupos, o CURE utiliza um conjunto de $1 < c < n$ objectos representativos, escolhidos

de forma a que o tamanho do grupo e respectivo formato possam ser caracterizados por apenas esses objectos. Assim, no processo de agrupamento aglomerativo, a distância entre dois grupos é calculada considerando apenas os objectos representativos mais próximos entre cada grupo.

Outro exemplo de um algoritmo de agrupamento hierárquico é o Chameleon [47]. Este algoritmo constrói inicialmente um grafo dos K-vizinhos-mais-próximos, ou seja, um grafo que tem um vértice por cada objecto x_i e cada vértice está ligado com um arco a K vértices, correspondendo aos K objectos mais próximos de x_i. Em seguida, é efectuado o particionamento do grafo dos K-vizinhos-mais-próximos, obtendo-se vários subgrupos de objectos. Finalmente, estes subgrupos são aglomerados sucessivamente até que seja encontrado um critério de paragem. A escolha dos grupos a fundir em cada iteração baseia-se numa função-objectivo que considera tanto a distância entre os objectos do mesmo grupo como a distância entre os objectos de grupos diferentes.

Abordagens baseadas em densidade. As abordagens acima descritas agrupam os objectos de dados baseando-se em medidas de distância. No entanto, existem outros algoritmos de agrupamento que se baseiam também na noção de densidade e têm como intuito encontrar regiões densas no espaço dos dados e fazer a correspondência entre os objectos que se encontram em cada região densa e os grupos do agrupamento de dados. A ideia principal destes algoritmos de agrupamento de dados consiste em fazer crescer um grupo enquanto a densidade da região que este engloba for superior a um determinado limiar.

O DBSCAN [30] é um algoritmo de agrupamento que segue uma abordagem baseada em densidade. Inicialmente, é escolhido ao acaso um objecto x_i, que tenha $MinPts$ objectos a uma distância não superior a Eps, sendo Eps o raio que define a vizinhança de x_i. De seguida, é formado um grupo encontrando todos os objectos que se encontrem a uma distância não superior a Eps de x_i. A etapa seguinte passa por alargar o grupo, adicionando todos os objectos que se encontrem a uma distância igual ou inferior a Eps de qualquer objecto já incluído no grupo (desde que a densidade mínima seja satisfeita). Estes passos repetem-se até que não seja possível adicionar objectos ao grupo. No final, a cada objecto do grupo é associado o rótulo correspondente. Em seguida, outro objecto não rotulado é seleccionado, sendo o processo acima descrito repetido para o novo objecto. O algoritmo termina quando não for possível formar mais nenhum grupo.

Abordagens baseadas em grelha. Os algoritmos de agrupamento baseados em grelha dividem os espaço dos dados em células, formando uma estrutura em forma de grelha. Cada atributo é dividido em vários intervalos, sendo cada objecto colocado na célula em que os intervalos associados incluem os valores dos atributos do objecto. No processo de agrupamento de dados apenas são consideradas as células da grelha e alguma informação estatística (por exemplo, o número de objectos e a densidade em cada célula). A vantagem dos algoritmos desta abordagem consiste na rapidez de processamento, pois, ao contrário

dos algoritmos de agrupamento de dados supramencionados, o custo de processamento é independente do número de objectos, sendo apenas dependente do número de células que a grelha contém, que é normalmente muito inferior ao número de objectos do conjunto de dados.

O algoritmo de agrupamento de dados STING (*STatistical INformation Grid*)[88] é um exemplo de um algoritmo baseado em grelha e funciona como explicado resumidamente de seguida. Inicialmente, constrói-se uma grelha para representar espacialmente todos os objectos do conjunto de dados, definindo hipercubos com um tamanho de aresta $TamAresta$ especificado pelo utilizador, isto é, cada atributo vai ser dividido em intervalos de tamanho $TamAresta$. Em seguida, cada objecto é atribuído à célula que engloba os valores dos seus atributos e, no final, a cada célula será adicionada informação relativa à sua posição no espaço, número dos objectos que possui e a descrição dos objectos que contém. As células que não contém qualquer objecto são automaticamente descartadas. Após a construção da grelha, os grupos de dados são formados pelas regiões densas da grelha, isto é, pelo conjunto de células vizinhas cuja densidade é igual ou superior a um limiar de densidade. Este passo é bastante semelhante à forma com que o algoritmo DBSCAN descobre os seus grupos, mas em vez de se procurar uma região densa na vizinhança de objectos, procura-se uma região densa na vizinhança de células na grelha.

O CLIQUE [1] (*CLustering In QUEst*) é também um algoritmo de agrupamento baseado em grelha. A construção da grelha difere ligeiramente do algoritmo STING, pois cada atributo é dividido num número predefinido de intervalos, em vez de ser indicado o tamanho de cada intervalo. A ideia do CLIQUE é bastante diferente da do algoritmo anterior, pois em vez de encontrar regiões densas que englobem todos os atributos de dados, o CLIQUE tem como objectivo descobrir correlações interessantes entre os objectos em sub-espaços de subconjuntos de atributos de dados.

Abordagens baseadas em modelos. Os algoritmos de agrupamento baseados em modelos têm como objectivo ajustar um modelo matemático ao conjunto de dados.

No agrupamento probabilístico assume-se que os objectos de dados foram gerados a partir de um modelo de mistura, existindo uma distribuição de probabilidade associada a cada grupo de dados. Assim, após se assumir uma determinada distribuição para os grupos de dados, geralmente a distribuição gaussiana, o problema do agrupamento de dados resume-se à estimação dos parâmetros que definem a função de densidade de probabilidade para cada um dos grupos. O algoritmo EM [24] (*Expectation-Maximization*) é um algoritmo muito usado para o efeito.

Outros exemplos de algoritmos baseados em modelos são o COBWEB [33] e os Mapas de Características Auto-Organizáveis [53] (*Self Organizing Maps* - SOM). O COBWEB é um algoritmo de agrupamento conceptual e incremental bastante popular em agrupamento de

dados categóricos. Em vez de atribuir um rótulo a cada objecto, este algoritmo define incrementalmente uma estrutura hierárquica, representante das relações entre os objectos de dados. A estrutura criada é semelhante a uma árvore de decisão, em que, cada nó da árvore corresponde a um conceito. A construção da árvore de decisão é realizada submetendo individualmente cada objecto ao algoritmo COBWEB, sendo então a hierarquia de conceitos (a árvore de decisão) actualizada tendo em conta uma medida de utilidade categórica. Um Mapa de Características Auto-Organizáveis consiste numa rede neuronal artificial, treinada de forma não supervisionada, que se organiza dinamicamente, respondendo aos estímulos de entrada, isto é, aos valores dos atributos dos objectos que são submetidos à rede. Este tipo de rede neuronal artificial é composto por duas camadas, uma de entrada e outra de saída. A camada de entrada da rede consiste num vector d-dimensional, em que d corresponde ao número de atributos do conjunto de dados. Cada objecto é submetido iterativamente à rede neuronal atribuindo os valores dos atributos dos objectos à camada de entrada. A camada de saída é composta por vários neurónios, dispostos numa estrutura de grelha, em que cada neurónio é caracterizado por um vector d-dimensional. Os neurónios da camada de saída competem pelos objectos, vencendo o neurónio cujo vector é mais similar ao objecto submetido. Em seguida, o vector do neurónio vencedor e os vectores dos neurónios seus vizinhos são actualizados para ficarem mais próximos do objecto submetido (segundo a distância euclidiana). Após todos os objectos de dados terem sido submetidos várias vezes ao treino da rede, espera-se que grupos de neurónios vizinhos vençam na competição por objectos semelhantes e que objectos dissemelhantes sejam associados a neurónios que não possuam qualquer relação de vizinhança.

2.4 Aprendizagem Semi-Supervisionada e Agrupamento de Dados com Restrições

Na aprendizagem semi-supervisionada pressupõem-se que nem todos os objectos pertencentes ao conjunto têm associado uma classe ou grupo. Na realidade, a grande maioria dos objectos de dados não têm atribuídos quaisquer valores para o atributo alvo, já que, o custo dessa atribuição pode ser bastante elevado, especialmente se a intervenção humana for necessária.

2.4.1 Classificação Semi-Supervisionada

Ao contrário da classificação supervisionada, em que para cada objecto do conjunto de treino a respectiva classe é conhecida, na aprendizagem semi-supervisionada conhece-se a classe de apenas alguns objectos. Para se compreender a aplicabilidade da classificação semi-supervisionada apresenta-se o exemplo seguinte. Imagine-se que se pretende construir um classificador usando um determinado conjunto de dados, cuja cardinalidade é bastante elevada. As classes a que os objectos podem pertencer são conhecidas, mas não existe informação sobre a classe a que

cada objecto pertence. O primeiro passo seria classificar cada um desses objectos para em seguida treinar um algoritmo de classificação tradicional. Se para essa classificação inicial fosse necessária a intervenção humana (por exemplo, o reconhecimento de objectos em imagens), esta seria extremamente morosa ou impraticável. Já classificar apenas alguns objectos representativos de cada classe seria viável. Assim, uma possibilidade passaria por classificar um número de objectos razoável desse conjunto de dados para em seguida treinar os algoritmos de aprendizagem supervisionada, descartando do conjunto de dados os objectos que não foram analisados. Esta opção pode originar que informação valiosa contida nos dados descartados não seja considerada no processo de aprendizagem. O objectivo da classificação semi-supervisionada é permitir que os objectos do conjunto de dados que não puderam ser classificados sejam úteis para a construção do classificador, aumentando se possível a qualidade da classificação de novos objectos.

Na literatura existem várias abordagens para a classificação semi-supervisionada de dados. Os modelos generativos são provavelmente os métodos mais antigos de aprendizagem semi-supervisionada [95]. Estes métodos assumem um modelo $p(x, l) = p(l)p(x|l)$, em que $p(x|l)$ é uma distribuição de mistura identificável, por exemplo, Gaussiana. O uso de objectos não rotulados pode melhorar a estimação dos parâmetros do modelo, melhorando assim a qualidade do classificador. A figura 2.2 ilustra esta ideia através de um problema com duas classes. A figura 2.2 a) representa a posição no espaço dos objectos rotulados, sendo os objectos da primeira classe representados por $\circ$ e os da segunda classe por $+$. Na figura 2.2 b) são acrescentados objectos não rotulados, sendo estes representados pelas áreas cinzentas. O modelo obtido com apenas o uso de objectos rotulados é ilustrado na figura 2.2 c). Como se pode verificar, o modelo apresentado ajusta-se demasiadamente aos objectos rotulados. Na figura 2.2 d), tal já não acontece, visto que o modelo se ajustou quer aos objectos rotulados quer aos não rotulados, aumentando a capacidade de generalização do modelo. O trabalho de Nigam [67] em classificação de texto é um bom exemplo do sucesso da classificação semi-supervisionada usando modelos generativos.

Outra abordagem bastante usada na classificação semi-supervisionada, devido à sua simplicidade, é o *Auto-Treino* (*Self-Training*) [60]. Inicialmente, é treinado um classificador usando todos os objectos rotulados, classificando-se em seguida todos os objectos não rotulados. Posteriormente, é avaliada a confiança associada à atribuição da classe a cada objecto. Os objectos que atinjam um determinado limiar de confiança são adicionados ao conjunto de dados rotulados, repetindo-se este processo um número especificado de vezes. Blum e Mitchel [13] propuseram o algoritmo de *Co-Treino* (*Co-Training*), que tem a particularidade de pressupor que o conjunto de atributos de dados pode ser dividido em dois subconjuntos, sendo cada subconjunto suficiente para treinar um bom classificador. Inicialmente, são treinados dois classificadores, f_1 e f_2, usando apenas objectos de dados rotulados, um para cada subconjunto de atributos, sendo alguns dos objectos não rotulados classificados, quer usando f_1 quer f_2, sendo posteriormente adicionados ao conjunto de objectos rotulados. Este processo é repetido até que seja encontrado um critério de paragem. A assumpção de que o conjunto de atributos pode ser dividido em dois

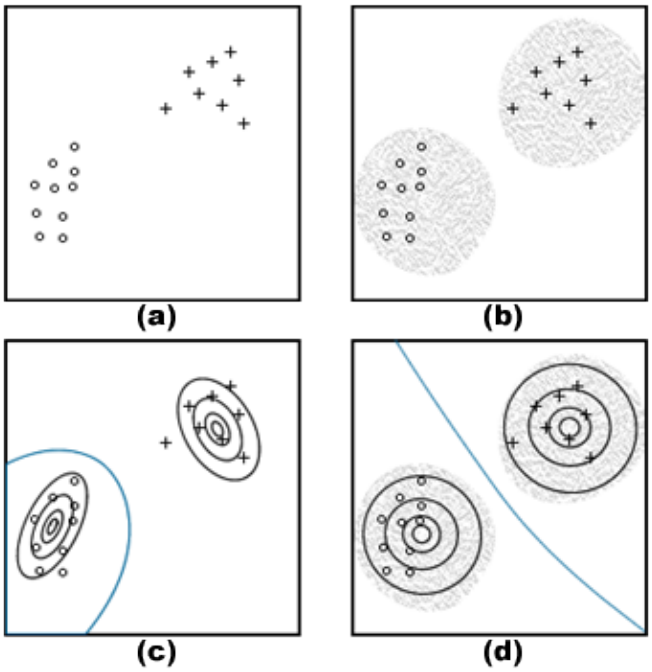

Figura 2.2: Exemplo da utilização de dados não rotulados para melhorar a qualidade de um modelo de classificação - (a) dados rotulados; (b) todos os objectos de dados disponíveis; (c) modelo aprendido usando apenas dados rotulados; (d) modelo aprendido usando todos os dados disponíveis.

subconjuntos condicionalmente independentes não é irreal, como ilustra a figura 2.3. Na classificação de imagens na *Web*, os atributos das imagens podem ser divididos em dois subconjuntos: o primeiro consiste nos atributos da própria imagem (cores, formas, etc.) e o segundo consiste no hipertexto que envolve essa imagem, que pode ser bastante útil na classificação da imagem.

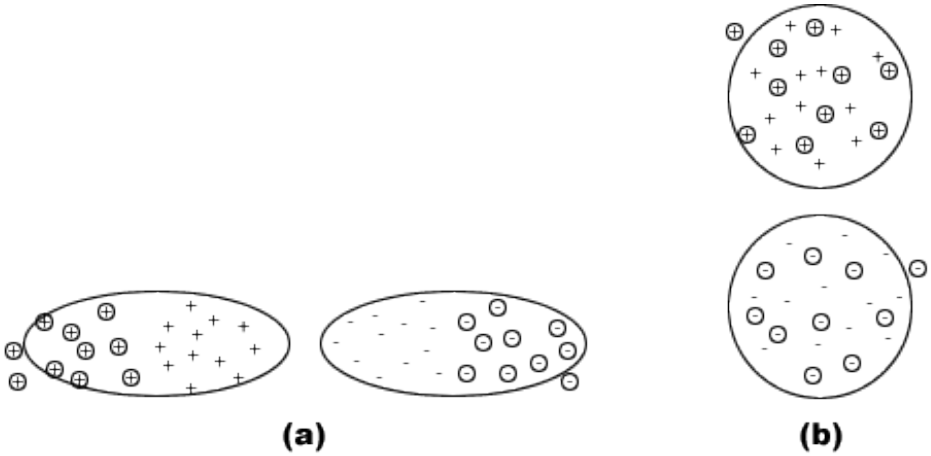

Figura 2.3: Exemplo de dois subconjuntos de atributos condicionalmente independentes - (a) Vista do conjunto de dados usando o primeiro subconjunto de dados; (b) Vista do conjunto de dados usando o segundo subconjunto de dados. Os símbolos + e − representam a classe objectos, sendo apenas conhecidas, para classificação, as classes dos objectos rodeados por uma circunferência

Um último exemplo de classificação semi-supervisionada, a máquinas de suporte vectorial semi-supervisionada (S^3VM - *Semi-Supervised Support Vector Machine*) foi introduzida por Bennett e Demiriz [10]. A sua ideia principal consiste em definir planos de decisão afastados de

regiões densas, populadas não só por objectos cuja classe é conhecida, mas também por objectos não rotulados, tal como ilustrado na figura 2.4.

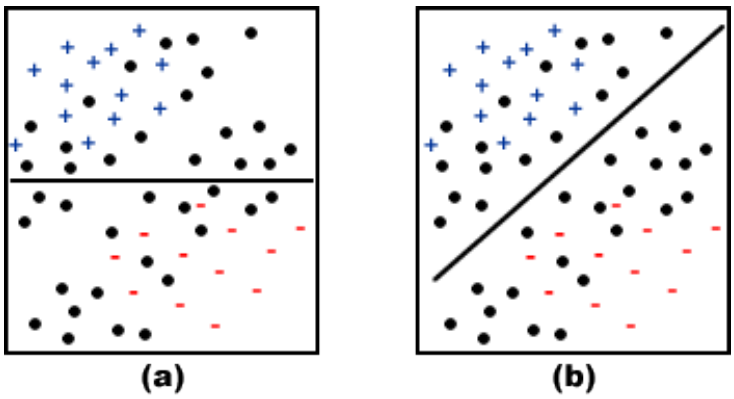

Figura 2.4: Máquina de Suporte Vectorial Semi-supervisionada - (a) Plano de decisão usando apenas objectos rotulados; (b) Plano de decisão usando objectos rotulados e não rotulados.

2.4.2 Agrupamento de Dados com Restrições

Uma área em grande desenvolvimento e bastante promissora é o agrupamento de dados com restrições. O agrupamento de dados com restrições permite que conhecimento *a priori* sobre o conjunto de dados possa ser incorporado no processo de agrupamento de dados. Esse conhecimento é representado na forma de restrições que expressam preferências, limitações e/ou condições que se pretendem impor no agrupamento dos dados, para que este seja útil e adequado a tarefas ou interesses específicos. O tema do agrupamento de dados com restrições será discutido em pormenor nos capítulos 3 e 4, sendo descritos os vários tipos de restrições que podem ser incorporados no agrupamento de dados e os principais algoritmos de agrupamento com restrições.

2.5 Sumário

Neste capítulo pretendeu-se introduzir alguns conceitos da aprendizagem automática com o objectivo de enquadrar o tema desta dissertação, o agrupamento de dados com restrições. Neste capítulo, são explicados os objectivos das aprendizagens supervisionada, não supervisionada e semi-supervisionada, sendo referidos alguns dos principais algoritmos de aprendizagem de cada área.

Capítulo 3

Agrupamento de Dados com Restrições

3.1 Introdução

O agrupamento de dados com restrições tem como objectivo utilizar o conhecimento sobre um determinado domínio na descoberta da estrutura do conjunto de dados. Esse conhecimento é representado na forma de restrições que expressam preferências, limitações e condições que o utilizador pretende impor. Espera-se assim que as soluções de agrupamento de dados se adeqúem da melhor forma a cada problema, pois o conhecimento representado através das restrições vai de encontro a resolução desse mesmo problema.

Na secção 3.2, são apresentados vários tipos de restrições, estando estes organizados pelo nível a que se encontram, desde as restrições mais gerais até às mais específicas, isto é, deste as restrições globais que se aplicam a todo o conjunto de dados até às restrições entre pares de objectos, que se encontram ao nível mais particular.

Na secção 3.3 apresenta-se sucintamente algumas abordagens para a aquisição inteligente de restrições.

3.2 Tipos de Restrições

O conhecimento de domínio do utilizador sobre um problema pode ser incluído no agrupamento de dados com o uso de restrições. As restrições podem ser categorizadas pelo nível a que se encontram. Ao nível mais alto situam-se as *restrições globais* que se aplicam a todo o conjunto de dados. As *restrições ao nível dos grupos* e as *restrições ao nível dos atributos* encontram-se no nível intermédio de especificidade. Por fim, as *restrições ao nível dos objectos de dados* estão no nível mais baixo. Os vários tipos de restrições são apresentados nas próximas subsecções.

3.2.1 Restrições Globais

Por vezes existem restrições que se pretendem aplicar a um determinado conjunto de dados $\mathcal{X}$ como um todo. Essas restrições são designadas por *restrições globais* e podem ter a forma de relações de vizinhança ou outro tipo de relações mais gerais entre os objectos de dados. De seguida, são apresentados vários métodos para incorporar restrições globais.

3.2.1.1 Obstáculos como Restrições

Considere-se uma empresa de telecomunicações que pretende determinar quais os melhores locais para colocar postos públicos de telefone numa determinada região. Para estimular o uso dos postos públicos de telefone, estes deverão estar situados em zonas movimentadas e de fácil acesso. As pessoas podem ser representadas por um conjunto de objectos de dados cujos atributos indicam a sua localização. Com a aplicação de um algoritmo de agrupamento de dados tradicional a esse conjunto de dados, os grupos obtidos representam regiões de elevada densidade populacional pelo que seria de esperar que os postos públicos de telefone fossem colocados nos centros dessas regiões. No entanto, os centros dos grupos podem situar-se em zonas em que a construção dos postos públicos é impossível, como é caso de, estradas, edifícios, zonas protegidas, etc. Mesmo que os centros dos grupos se localizem em locais possíveis, a simples distância euclidiana pode não ser adequada para o problema. Considere-se que o centro de um grupo se situa próximo da margem de um rio. Uma pessoa que more na outra margem do rio pode ser incluída nesse grupo, apesar de poder ter de percorrer uma grande distância para se deslocar a aquele posto público de telefone, já que poderá não existir nenhuma ponte próxima.

Um exemplo de um algoritmo de agrupamento de dados com obstáculos é o COE-CLARANS [42] baseado no algoritmo de agrupamento de dados CLARANS [66]. O algoritmo constrói um grafo, onde representa os objectos de dados e os obstáculos, e utiliza técnicas de geometria computacional para calcular as distâncias mais curtas entre os objectos, tendo em conta que os obstáculos têm de ser contornados. Geralmente a quantidade de dados espaciais é bastante elevada, pelo que são realizados vários pré-processamentos e optimizações para reduzir o custo computacional do algoritmo.

3.2.1.2 Informação de Vizinhança

Os algoritmos de agrupamento de dados são por vezes aplicados a conjuntos de dados em que os objectos de dados se encontram relacionados através de informação estrutural ou de vizinhança. Wagstaff [87] explica várias formas de incorporar este tipo de informação usando o exemplo de uma imagem constituída por vários *pixels* organizados em posições bidimensionais. A figura 3.1 (a) representa uma imagem com 25 *pixels* com três regiões distintas. As relações de vizinhança entre os *pixels* são representadas pelas ligações ilustradas na figura 3.1 (b). Finalmente, a figura 3.1 (c) exibe o agrupamento dos *pixels* em três grupos, considerando quer as relações de vizinhança, quer os atributos de cada *pixel* (neste caso, as posições e cor de cada *pixel*)

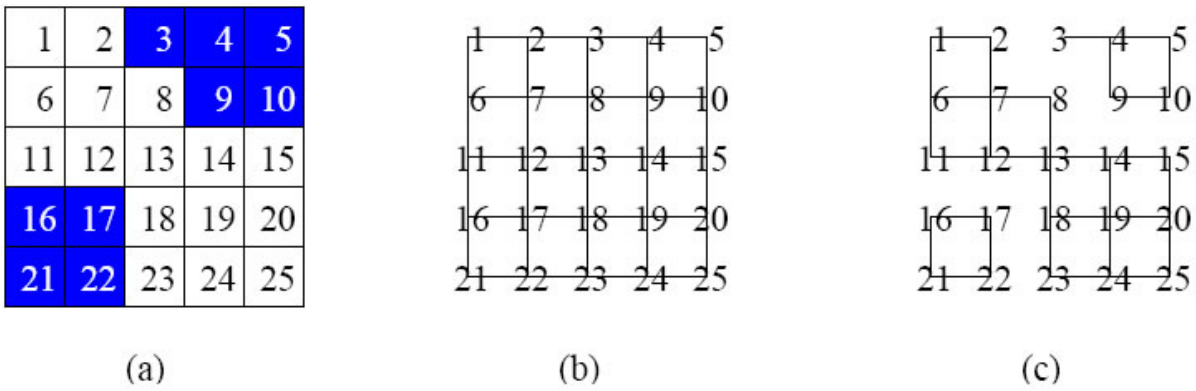

Figura 3.1: Segmentação de imagem - (a) imagem com os *pixels* numerados; (b) relações de vizinhança; (c) Agrupamento dos *pixels* em 3 grupos, sujeitos à relação $V(x_i)$ (retirado de [87]).

Continuidade espacial. Na segmentação de uma imagem, isto é, no agrupamento de *pixels*, é mais provável que dois *pixels* x_i e x_j adjacentes pertençam ao mesmo grupo que dois *pixels* muito afastados. Por esta razão é natural que se pretenda aperfeiçoar a continuidade dos grupos de *pixels*. É útil definir uma relação de vizinhança para cada um dos *pixels*, atribuindo a um *pixel* x_i um conjunto de vizinhos $V(x_i)$. Cada conjunto de vizinhos é geralmente composto pelos quatro *pixels* imediatamente adjacentes (equação 3.1), ou seja, os vizinhos norte, sul, oeste e este, ou pelo conjunto dos oito vizinhos mais próximos que englobam também os *pixels* das diagonais (equação 3.2). Assim, a vizinhança de um *pixel* x_i pode ser exprimida por uma das seguintes equações:

$$V(x_i) = \{x_j | d_E(x_i, x_j) \leq 1\} \tag{3.1}$$

$$V(x_i) = \{x_j | d_E(x_i, x_j) < 2\} \tag{3.2}$$

em que d_E representa a distância euclidiana referente à posição espacial dos *pixels* na imagem. Note-se que esta distância é independente da medida de distância $d(x_i, x_j)$ usada para efectuar o agrupamento de dados, neste caso, a segmentação da imagem. As equações 3.1 e 3.2 são relativas às vizinhanças de quatro e oito *pixels*, respectivamente. De seguida são apresentados cinco métodos para a incorporação de informação de vizinhança.

1. **Adição de Atributos Posicionais.** Uma das formas mais simples de incorporar a noção de posição no espaço dos objectos de um conjunto de dados $\mathcal{X}$, consiste em adicionar atributos extra que representem a posição dos objectos no conjunto de dados. No caso da segmentação de uma imagem bidimensional, os atributos extra podem ser a *linha* e *coluna* de cada *pixel*. Assim, a função de distância entre dois *pixels* $d(x_i, x_j)$ atribui naturalmente uma maior dissemelhança a *pixels* afastados e uma maior semelhança a *pixels* próximos. A simplicidade deste método constitui uma vantagem obvia, já que, não requer alterações do algoritmo de agrupamento de

dados. Contudo, este método acrescenta apenas uma preferência de continuidade espacial dos grupos de *pixels*, não impondo um requisito espacial.

2. **Duplicação de Vizinhos como Atributos.** Outro método para incorporar a noção de posição no espaço no agrupamento de dados resume-se a aumentar os atributos de um conjunto de dados $\mathcal{X}$ com os atributos respeitantes a cada vizinho [43]. Apesar deste método ser bastante simples é muitas vezes impraticável devido à elevada dimensionalidade que o conjunto de dados pode tomar. Basicamente, cada objecto de dados (*pixel* neste caso) x_i recebe cópias dos atributos dos seus objectos de dados vizinhos aumentando o número de atributos do conjunto de dados e afectando negativamente o desempenho dos algoritmos de agrupamento de dados. Actualmente, com o volume de informação a crescer exponencialmente (seguindo o exemplo da segmentação de imagens, actualmente as máquinas fotográficas digitais possuem resoluções na ordem das dezenas de *megapixel*) este método pode tornar o agrupamento de dados impraticável. Este método, tal como o anterior, não impõe um requisito espacial.

3. **Modificação do Cálculo da Distância.** Os dois métodos apresentados anteriormente alteram o conjunto de dados $\mathcal{X}$. Outra forma de utilizar a informação da vizinhança consiste na modificação de como a distância entre dois objectos de dados x_i e x_j é calculada, não havendo assim necessidade de alterar o conjunto de dados $\mathcal{X}$ [68]. A principal desvantagem deste método é a especialização do algoritmo de agrupamento de dados a problemas que tenham acesso à mesma informação de domínio, já que a forma como a distância entre objectos de dados é calculada é dependente dessa informação.

4. **Modificação da Função Objectivo.** Outra abordagem para usar a informação de vizinhança no agrupamento de dados resume-se à modificação da função-objectivo do próprio algoritmo de agrupamento de dados. Geralmente, essa alteração é realizada de forma a que a nova função-objectivo seja uma soma ponderada da função-objectivo original do algoritmo de agrupamento de dados e de uma outra função relativa à informação da vizinhança, em que as ponderações de ambas as funções reflectem a importância que o utilizador dá à optimização da função-objectivo original e à informação da vizinhança [78].

5. **Restrições Directas.** Finalmente, a informação de vizinhança pode ser incluída no agrupamento de dados restringindo o conjunto de agrupamentos permitidos. Neste caso, os grupos resultantes do agrupamento de dados têm de cumprir todas as restrições impostas pelo utilizador. Nos algoritmos de agrupamento de dados hierárquicos aglomerativos, as restrições directas podem ser impostas restringindo os pares de grupos que se considera fundir [83]. No caso dos algoritmos de agrupamento de partição,

as restrições directas podem ser impostas modificando o passo de atribuição dos objectos de dados aos grupos cujos centros sejam mais próximos, considerando a atribuição de um objecto de dados a apenas grupos em que as restrições não sejam violadas [86].

3.2.2 Restrições ao Nível dos Grupos

Por vezes, existe informação que se pretende aplicar aos grupos (C_i) de objectos de dados individualmente e não à totalidade do conjunto de dados $\mathcal{X}$. Esta informação é representada usando restrições ao nível dos grupos e têm como objectivo restringir a forma, o tamanho, a variância ou outras características dos grupos. As restrições ao nível dos grupos mais frequentemente usadas são do tipo capacidade mínima e máxima, isto é, restrições que limitam o número mínimo e máximo de objectos de dados que cada grupo pode conter.

Considere-se o exemplo da empresa de telecomunicações apresentado na secção 3.2.1.1. Esta pretende colocar K postos públicos de telefone, tal que, a distância a percorrer por cada cliente até a um posto público (centro de grupo) seja minimizada e que cada posto (centro de grupo) se situe numa região com um limiar mínimo de densidade populacional. Neste caso, o algoritmo de agrupamento a utilizar teria de suportar restrições ao nível do número mínimo de objectos necessários para formar cada grupo. Noutras aplicações, pode ser útil usar outras restrições ao nível do grupo para além do número (mínimo ou máximo) de objectos, como é o caso das as variâncias mínima e máxima, e do raio mínimo e máximo.

3.2.2.1 Restrições de Capacidade Mínima

Um problema com o algoritmo de agrupamento de dados K-Médias consiste na obtenção de grupos vazios ou com um número reduzido de objectos, que ocorre frequentemente na sua utilização em conjuntos de dados de elevada dimensionalidade e especialmente quando o número de grupos pretendidos (K) é elevado. Para solucionar este problema, Bradley *et al.* [15] propuseram uma modificação no algoritmo K-Médias que determina que o número de objectos de dados em cada grupo não pode ser inferior a um valor especificado pelo utilizador e a atribuição dos objectos aos K grupos é visto como um problema de fluxo de custo mínimo. Esta especificação é uma restrição da capacidade mínima dos grupos e torna o algoritmo K-Médias menos sensível a mínimos locais. Uma proposta de Tung *et al.* [82] para resolver a restrição da significância (número mínimo de objectos) do grupo consiste em começar num agrupamento de dados qualquer, desde que este satisfaça as restrições de significância, e em seguida movimentar alguns objectos de dados entre os vários grupos de forma a minimizar a sua dispersão, mantendo no entanto todas as restrições satisfeitas. Ge *et al.* [38] desenvolveram um algoritmo que, para além das restrições de significância, suporta restrições relativas às variâncias (minímas e máximas) dos grupos e procura no conjunto de dados um número arbitrário de grupos que sejam compactos e balanceados. Este algoritmo é baseado numa estrutura de dados, denominada *CD-Tree*,

que armazena os objectos de dados nas folhas de uma árvore em que cada folha satisfaz aproximadamente as restrições impostas e minimiza a função-objectivo (a soma dos quadrados das distâncias entre os objectos e o centro do grupo a que estão associados).

3.2.2.2 Restrições de Capacidade Máxima

Em alguns problemas pode também ser útil definir a capacidade máxima, isto é, o número máximo de objectos em cada grupo de dados. Voltando ao exemplo da empresa de telecomunicações da secção 3.2.1.1, a empresa pode pretender colocar os postos públicos de telefone em determinados locais de forma a que cada posto seja utilizado por um número limitado de utilizadores e fazendo com que o tempo de espera dos seus clientes para a utilização do mesmo não seja muito elevado. As restrições de capacidade máxima são também utilizadas frequentemente na análise de localizações para infra-estruturas [75].

3.2.3 Restrições ao Nível dos Atributos

O conhecimento de domínio que um utilizador possui para um determinado problema pode permitir a criação de restrições derivadas directamente dos atributos do conjunto de dados em análise. A partir dos valores dos atributos de dados, o utilizador pode definir heurísticas que influenciem positivamente o agrupamentos dos objectos de dados. Por exemplo, o utilizador pode pretender que todos os objectos de dados que possuam o mesmo valor para um determinado atributo sejam agrupados no mesmo grupo. Uma das formas de o fazer consiste na conversão desta restrição ao nível dos atributos para várias restrições de relações entre pares de objectos de dados. As relações entre pares de objectos de dados serão abordadas um pouco mais à frente no tópico 3.2.4.2.

3.2.4 Restrições ao Nível dos Objectos

Ao nível mais particular do conhecimento de domínio encontram-se as restrições ao nível dos objectos. Neste tipo de restrições, a pertença dos objectos de dados a grupos é restringida considerando informação relativa aos próprios objectos de dados, sendo impostas relações de pertença aos grupos entre pares de objectos de dados. As restrições ao nível dos objectos de dados podem assumir várias formas, tais como, rotulação parcial do conjunto de dados, restrições na colocação relativa entre pares de objectos ou informação resultante da interacção do utilizador em sistemas interactivos. De seguida são apresentados alguns exemplos de cada uma destas formas de restrições ao nível dos objectos de dados.

3.2.4.1 Rotulação Parcial

É bastante difícil obter um conjunto de dados em que todos os objectos de dados se encontram rotulados, já que, normalmente a atribuição das classes ou grupos aos objectos de dados é

efectuada com a intervenção humana, o que pode ser muito dispendioso quer a nível de tempo quer económico [95]. No entanto, a obtenção de apenas um subconjunto de dados rotulado é frequentemente exequível. Por este motivo e caso esta informação esteja disponível, a utilização desta no agrupamento de dados pode ser bastante vantajosa.

Votação Maioritária com base nos Objectos Rotulados. A abordagem de Demiriz *et al.* [23] realiza o agrupamento de dados usando um algoritmo genético que pretende minimizar uma função de custo que combina linearmente a dispersão dos objectos de dados aos seus grupos e a impureza do agrupamento de dados (usando o índice Gini[16]). Os rótulos existentes em alguns dos objectos de dados restringem a colocação dos objectos de dados nos grupos com base na medida de impureza. No processo de agrupamento cada grupo é rotulado usando votação maioritária tendo em conta os objectos rotulados que possui.

Grupos Sementeados com base em Rótulos. Em vez de utilizar os rótulos de alguns objectos de dados para determinar a identidade (classe) de cada grupo, Basu *et al.* [5] usa os objectos de dados rotulados para seleccionar de forma inteligente os centros iniciais de cada grupo de objectos de dados na inicialização do algoritmo K-Médias. Os centros iniciais dos grupos são determinados da seguinte forma: inicialmente os objectos de dados cujos rótulos são conhecidos são separados em K grupos, de forma a que cada grupo possua apenas objectos com o mesmo rótulo; em seguida, cada um dos K centros de grupos iniciais são obtidos calculando a média dos vectores que representam os objectos rotulados incluídos no grupo.

3.2.4.2 Relações entre Pares de Objectos de Dados

A forma mais generalista e flexível para representar o conhecimento de domínio usando informação relacional consiste na definição de um conjunto de relações entre pares de objectos de dados [87]. Com este tipo de representação é possível codificar várias das outras formas de conhecimento de domínio já apresentadas. As restrições globais que definem relações de vizinhança podem ser facilmente transformadas em relações entre pares de objectos de dados. O mesmo acontece relativamente às restrições ao nível dos atributos em que, geralmente, as heurísticas definidas pelo utilizador indicam se dois objectos de dados devem ou não ser agrupados conjuntamente. Apenas as restrições ao nível dos grupos podem ser dificilmente transformadas em relações entre pares de objectos, já que se focam principalmente nas capacidades mínimas e máximas de cada um dos grupos do agrupamento de dados a formar.

Nas relações entre pares de objectos de dados, as restrições mais frequentemente utilizadas são do tipo ligação obrigatória ou ligação proibida. Uma ligação obrigatória, $(x_i, x_j) \in Rest_{=}$, indica que dois objectos de dados x_i e x_j devem ser agrupados no mesmo grupo, enquanto que uma ligação proibida, $(x_i, x_j) \in Rest_{\neq}$, indica que os dois objectos de dados x_i e x_j não devem ser agrupados conjuntamente.

Estas restrições podem ser vistas como restrições inflexíveis, em que, o agrupamento de um conjunto de dados não pode violar nenhuma das restrições definidas em $Rest_{=}$ e $Rest_{\neq}$, ou como restrições flexíveis, em que, as restrições contidas em $Rest_{=}$ e $Rest_{\neq}$ indicam apenas as preferências do utilizador no agrupamento de dados, devendo portanto serem tidas em conta, não sendo no entanto invioláveis.

Restrições Inflexíveis. As restrições inflexíveis consistem em restrições entre pares de objectos declaradas pelo utilizador que têm que obrigatoriamente ser satisfeitas pelo algoritmo de agrupamento de dados. Assim, dois objectos de dados x_i e x_j relacionados com uma ligação obrigatória $(x_i, x_j) \in Rest_{=}$ têm obrigatoriamente de ficar colocados no mesmo grupo C_l na solução produzida pelo algoritmo de agrupamento, $(x_i, x_j) \in C_l$. Dois objectos de dados x_i e x_j relacionados com uma ligação proibida $(x_i, x_j) \in Rest_{\neq}$ têm obrigatoriamente de ser colocados em grupos diferentes, C_l e C_m, no agrupamento de dados retornado pelo algoritmo de agrupamento, $x_i \in C_l$, $x_j \in C_m$, tal que, $i \neq j$ e $l \neq m$.

Restrições Flexíveis. As restrições flexíveis podem ser vistas como preferências que o utilizador gostaria que fossem acatadas no agrupamento do conjunto de dados, não tendo no entanto de ser necessariamente satisfeitas. Estas preferências podem ser tratadas pelo algoritmo de agrupamento de dados com o mesmo grau de importância ou podem conter ponderações que indicam o grau com que o utilizador as quer ter satisfeitas na solução obtida pelo agrupamento de dados. Os valores das ponderações indicam assim a confiança que o utilizador tem nas relações entre pares de objectos de dados que define. Uma das formas para definir uma ponderação w_{ij} na relação entre dois objectos de dados x_i e x_j consiste na atribuição de valores no intervalo $[-1; 1]$, em que $w_{ij} > 0$ indica preferência em agrupar conjuntamente os dois objectos de dados e $w_{ij} < 0$ preferência em colocar os dois objectos de dados em grupos distintos [87]. Quanto maior a amplitude de w_{ij} (o seu valor absoluto) maior é a vontade com que o utilizador deseja que a relação seja satisfeita. Uma ponderação definida a 0, $w_{ij} = 0$, reflecte a inexistência de preferência do utilizador na relação entre os objectos de dados x_i e x_j. Existem várias outras formas de codificar o grau de confiança com que o utilizador determina as relações entre objectos e estas serão abordadas na secção 4.1, a propósito da discussão de vários algoritmos de agrupamento de dados que usam restrições.

3.2.4.3 Interactividade com o Utilizador

Este tipo de restrições é obtido através da interacção de um sistema de agrupamento de dados com o utilizador de forma iterativa. Em cada iteração, o sistema de agrupamento de dados produz uma solução de agrupamento e apresenta-a ao utilizador. Este avalia a solução apresentada e indica os erros que o sistema produziu, para que, nas iterações seguintes essa informação seja utilizada no sentido de evitar erros semelhantes. O trabalho de Cohn *et al.* [20] apresentado na secção 3.3.2 é um bom exemplo desta forma de obter restrições ao nível dos objectos de dados.

3.3 Aquisição de Restrições

A aprendizagem activa para a aquisição de restrições tem como objectivo ajudar o utilizador na definição de restrições sobre o conjunto de dados em análise, para que, essas restrições tenham um impacto positivo no processo de agrupamento de dados. Assim, espera-se que as restrições adquiridas sejam o mais informativas possíveis e que os algoritmos de agrupamento com restrições aproveitem essas restrições para obter partições de dados com qualidade superior.

3.3.1 Explorar e Consolidar

Basu *et al.* propuzeram um esquema de aprendizagem activa para adquirir relações de ligação obrigatória e proibida [7]. A abordagem consiste em dois passos: *Explorar* e *Consolidar*. No passo Explorar, o conjunto de dados é pesquisado com o intuito de encontrar conjuntos de vizinhança entre os pares de objectos, em que cada conjunto de vizinhança pertence a um grupo *natural* do conjunto de dados. Neste passo, o utilizador é consultado iterativamente sobre se dois objectos devem (ou não) ser atribuídos ao mesmo grupo. No passo Consolidar, a partir das consultas efectuadas ao utilizador pretende-se obter a melhor informação que possibilite uma boa estimação dos centros iniciais dos grupos do conjunto de dados.

3.3.2 Interacção com o Utilizador

Uma outra abordagem baseia-se na observação de que "é mais fácil criticar que construír" [20]. A proposta de Cohn *et al.* consiste em três etapas:

- Agrupar o conjunto de dados com um algoritmo de agrupamento não supervisionado.
- Analisar o resultado do agrupamento, em que o utilizador verifica em apenas alguns dos grupos se existem objectos mal agrupados. O utilizador dá informações ao sistema numa das seguinte formas:
  - o objecto não pertence ao grupo a que foi atribuído;
  - objecto deve ser movido para um determinado grupo;
  - dois determinados objectos devem ser atribuídos ao mesmo grupo;
  - e dois determinados objectos não podem ser atribuídos ao mesmo grupo.
- Após a solução ter sido criticada, a medida de distância do algoritmo de agrupamento é modificada para que as restrições impostas pelo utilizador sejam satisfeitas.

O processo supracitado repete-se enquanto o utilizador não estiver satisfeito com a solução apresentada.

3.4 Sumário

Neste capítulo apresentou-se uma visão geral dos vários tipos de restrições que podem ser incluídas no agrupamento de dados. O uso de restrições permite que o utilizador indique preferências, limitações e conhecimento de domínio no agrupamento de dados para que a solução obtida seja mais útil e vantajosa para os seus objectivos. No capítulo 4, *Algoritmos de Agrupamento de Dados com Restrições*, são apresentados vários algoritmos de agrupamento de dados que permitem a inclusão de restrições.

Capítulo 4

Algoritmos de Agrupamento de Dados com Restrições

4.1 Introdução

Durante a última década foram propostos bastantes algoritmos de agrupamento de dados com a capacidade de incorporar restrições, explorando várias ideias e tipos de restrições. A sua grande maioria incide nas restrições ao nível dos objectos de dados, mais precisamente na relações entre pares de objectos de dados (ligações obrigatórias/proibidas), pelo que os algoritmos apresentados neste capítulo focam principalmente este último tipo de restrições.

Nas próximas secções são apresentados alguns dos principais algoritmos de agrupamento de dados que usam restrições, estando estes divididos em cinco categorias: *Restrições Invioláveis* (secção 4.2), *Restrições na Forma de Rótulos* (secção 4.3), *Penalização de Violações de Restrições* (secção 4.4), *Edição de Distância* (secção 4.5) e *Modificação do Processo de Geração* (secção 4.6).

4.2 Restrições Invioláveis

Os algoritmos de agrupamento de dados com restrições descritos nesta secção têm a particularidade de garantir que a solução produzida satisfaz sempre todas as restrições especificadas pelo utilizador. Na subsecção 4.2.1 é apresentado o algoritmo de agrupamento de dados COP-COBWEB que trata apenas de conjuntos de dados cujos atributos são categóricos. Na subsecção 4.2.2 é apresentado o algoritmo COP-K-médias, podendo este ser aplicado a conjuntos de dados com atributos numéricos.

4.2.1 COP-COBWEB

O algoritmo de agrupamento de dados com restrições COP-COBWEB [85] (*COnstraint-Partitioning COBWEB*) é baseado no COBWEB [33], um algoritmo de agrupamento de dados incremental bastante conhecido. O COP-COBWEB emprega o conceito de *utilidade categórica* para produzir um agrupamento de dados que maximiza a dissimilaridade entre os grupos e similaridade dos objectos pertencentes ao mesmo grupo, definida pela seguinte equação:

$$\frac{\sum_{k=1}^{K} Pr(C_k) \sum_{i}^{d} \sum_{j}^{m} Pr(A_i = V_{ij}|C_k)^2 - \sum_i \sum_j Pr(A_i = V_{ij})^2}{K} \tag{4.1}$$

em que K é o número de grupos, C_k representa o k-ésimo grupo, A_i refere-se a um dos d atributos, V_{ij} é um dos m valores que A_i pode tomar e $Pr(\cdot)$ representa a probailidade de um evento.

O COBWEB tem quatro operadores primários, nomeadamente *Adicionar*, *Novo*, *Fundir* e *Dividir*, que representam as formas possíveis de incorporar um objecto de dados x_i no nível de topo da hierárquica corrente. Para cada objecto de dados x_i, cada um dos operadores é aplicado e o agrupamento resultante que obtiver maior *utilidade categórica* é seleccionado. O COBWEB aplica, recursivamente, os mesmos operadores aos filhos do grupo em que x_i foi inserido, para colocar o novo objecto ordenadamente nos níveis de maior profundidade da hierarquia, parando quando x_i alcançar um nó folha. O COP-COBWEB (algoritmo 4.1) retorna como agrupamento de dados, o nível de topo da hierarquia produzida pelo COBWEB e satisfaz todas as restrições impostas pelo utilizador. Para cada objecto de dados x_i submetido ao algoritmo, são inicialmente verificadas as restrições de *ligação obrigatória* (passo **Verificar**). Se existir alguma restrição de *ligação obrigatória* que indique que algum objecto x_i tem de ser associado ao mesmo grupo que outro objecto x_j, já existente num grupo C_k do agrupamento de dados corrente, o objecto x_i é incluído no grupo C_k. Caso contrário, os operadores **Novo**, **Adicionar** e **Fundir** são aplicados para se determinar em que grupo x_i vai ser incluído.

As restrições do tipo *ligação proibida* são verificadas nos passos **Adicionar** e **Fundir**. Quando se considera incluir um objecto x_i num grupo C_k, verifica-se se algum dos objectos x_j pertencentes a C_k tem uma relação de *ligação proibida* com x_i e, se existir, x_i não poderá ser incluído em C_k. O mesmo se passa na avaliação da fusão entre dois grupos. O operador **Dividir** é sempre avaliado, quer tenha sido encontrada ou não alguma restrição no passo **Verificar**. Este operador aplica recursivamente o COP-COBWEB a um subconjunto de dados que corresponde ao melhor grupo em que x_i foi colocado, segundo a *utilidade categórica*. Finalmente, o agrupamento de dados resultante é aquele em que a *utilidade categórica* é maximizada.

Algoritmo 4.1: COP-COBWEB

Entrada: $\mathcal{X}$ - Conjunto de dados, $Rest_{=}$ - Restrições de ligação obrigatória, $Rest_{\neq}$ - Restrições de ligação proibida.

Saída: P - Agrupamento de dados

Seja P um conjunto de grupos vazio, $P \leftarrow \{\}$;
para cada $x_i \in \mathcal{X}$ **faça**
  /*Considerar todas as formas de incorporar x_i em P;
  /*Verificar restrições de Ligação Obrigatória;
  se $\exists (x_i, x_j) \in Rest_{=}$ *em que* x_j *pertence a um grupo* $C_k \in P$ **então**
    $P_{ligObrig} \leftarrow (P - C_k) \bigcup \{C_k \bigcup \{x_i\}\}$;
  senão
    /*Adicionar x_i a um grupo;
    para cada $C_k \in P$ **faça**
      se $(x_i, x_j) \nexists \in Rest_{\neq}, \forall x_j \in C_k$ **então**
        $P_{adicionar_k} \leftarrow (P - C_k) \bigcup \{C_k \bigcup \{x_i\}\}$;
      fim
    fim
    /* Novo grupo com x_i ;
    $P_{novo} \leftarrow P \bigcup C$ em que $C = \{x_i\}$ é um novo grupo ;
    /*Fundir os dois grupos C_{max_1} e C_{max_2} com maior UC pertencentes às melhores partições obtidas no passo Adicionar ;
    se $\nexists (x_l, x_m) \in Rest_{\neq}, x_l \in C_{max_1}$ *e* $x_m \in C_{max_2}$ **então**
      $P_{fundir} \leftarrow ((P - C_{max_1}) - C_{max_2}) \bigcup \{C_{max_1} \bigcup C_{max_2} \bigcup \{x_i\}\}$;
    fim
  fim
  /*Dividir o grupo C_{max} correspondente ao melhor grupo para x_i, segundo UC, resultante das partições obtidas nos passos Verificar e Adicionar;
  $P_{dividir} \leftarrow (P - C_{max}) \bigcup COP - COBWEB(C_{max} \bigcup \{x_i\}, Rest_{=}, Rest_{\neq})$
  Atribuir
  $K \leftarrow \arg\max_k UC(P_k)$ para todo $k \in \{ligObrig, adicionar_j, novo, fundir, dividir\}$;
  Actualizar $P \leftarrow P_K$;
fim
Devolver P;

4.2.2 COP-*K*-médias

O K-médias [58] é um dos algoritmos de agrupamento mais conhecidos e usados. Este algoritmo de partição divide um conjunto de dados $\mathcal{X}$ em K grupos usando um processo iterativo bastante simples. Inicialmente, K centros de grupos são escolhidos aleatoriamente e são iterativamente aperfeiçoados em dois passos:

1. Cada objecto x_i é atribuído ao grupo cujo centro é mais próximo.

2. Cada centro de grupo $\overline{x}_k$ é recalculado como a média dos objectos que o constituem.

Os dois passos anteriores são repetidos iterativamente até que não existam mudanças no grupo atribuído aos objectos. O algoritmo COP-K-médias [86] (*COnstraint-Partitioning K-means*) é uma modificação do K-médias para que este possa suportar restrições de *ligação obrigatória*

e de *ligação proibida* entre pares de objectos de dados. Nesse sentido, a grande alteração no K-médias é realizada no passo de atribuição de cada objecto de dados x_i ao grupo C_k mais próximo. Antes de se atribuir x_i ao grupo mais próximo é realizado um passo de verificação de violações das restrições (algoritmo 4.2). Desta forma, o objecto x_i vai ser incluído no grupo mais próximo em que todas as restrições sejam satisfeitas. Caso não exista nenhum grupo que satisfaça todas as restrições que envolvem x_i, o algoritmo aborta e retorna {}.

Algoritmo 4.2: Violação de Restrições

Entrada: x_i - objecto de dados, C - Grupo, $Rest_=$ - Restrições de ligação obrigatória, $Rest_{\neq}$ - Restrições de ligação proibida, K - Número de grupos pretendido.

```
para cada (x_i, x_=) ∈ Rest_= faça
    se x_= ∉ C então
        retorna verdadeiro
    fim
fim
para cada (x_i, x_≠) ∈ Rest_≠ faça
    se x_≠ ∈ C então
        retorna verdadeiro
    fim
fim
retorna falso
```

Algoritmo 4.3: COP-K-médias

Entrada: $\mathcal{X}$ - Conjunto de dados, $Rest_=$ - Restrições de ligação obrigatória, $Rest_{\neq}$ - Restrições de ligação proibida, K - Número de grupos pretendido.

Saída: P - Agrupamento de dados

```
Obter aleatoriamente K centros de grupos, C_1, ··· , C_K;
repita
    Atribuir x_i ao grupo C_k com o centro mais próximo em que Violação de
    Restrições(x_i, C_k, Rest_=, Rest_≠ ) = falso ;
    se x_i não é atribuído a qualquer grupo então
        retorna {};
    fim
    para cada k ← 1 até K faça
        Actualizar o centro do grupo C_k com a média de todos os objectos x_i ∈ C_k
    fim
até não existir alteração dos grupos ;
Devolver P = {C_1, ··· , C_K};
```

O algoritmo de agrupamento de dados COP-K-médias é descrito no algoritmo 4.3, mostrando as alterações efectuadas ao K-médias para incorporar restrições entre pares de objectos de dados.

4.3 Restrições na Forma de Rótulos

Nesta categoria encontram-se os algoritmos de agrupamento cujas restrições são expressas através de rótulos. Neste caso, são conhecidos os grupos a que pertencem alguns dos objectos de dados e os algoritmos de agrupamento de dados com restrições usam essa informação para aumentar a qualidade de agrupamento de dados. Na próxima secção são apresentados dois exemplos de algoritmos de agrupamento de dados com restrições que pertencem a esta categoria, os algoritmos K-médias *semeado* e *restringido* [5].

4.3.1 K-médias Semeado e Restringido

Uma das formas de incorporar restrições na forma de rótulos no agrupamento de dados consiste na utilização dos objectos de dados rotulados para a geração de *sementes* de grupos e usá-las para inicializar o algoritmo de agrupamento de dados. A motivação desta abordagem é simples: se as sementes dos grupos iniciais forem apropriadas, a pesquisa do algoritmo de agrupamento é dirigida para uma boa região do espaço de pesquisa, o que simultaneamente reduz significativamente as hipóteses do algoritmo de agrupamento terminar a sua pesquisa em óptimos locais de pouca qualidade, permitindo que o agrupamento de dados obtido seja concordante com os rótulos definidos pelo utilizador. Basu *et al.* [5] exploram esta abordagem propondo duas variantes do algoritmo de agrupamento de dados K-médias: o K-médias semeado (*Seeded-KMeans*) e o K-médias restringido (*Constrained-KMeans*). Considere-se $\mathcal{X} = \{x_1, \cdots, x_n\}$ um conjunto de n objectos de dados e $S \subseteq \mathcal{X}$ o conjunto de sementes que é composto por todos os objectos rotulados. É também assumido que S está dividido em K partições, $\{S_1, \cdots, S_K\}$, tal que exista um subconjunto S_j, com pelo menos um objecto de dados, para cada um dos K grupos *naturais*. No K-médias semeado, o conjunto de sementes é apenas utilizado para definir os centros dos grupos iniciais do K-médias, em vez de estes serem definidos aleatoriamente. Nesse sentido, o centro $\overline{x}_j$ do j-ésimo grupo inicial, C_j, é obtido pela média dos objectos que constituem o j-ésimo subconjunto de sementes, S_j. O resto do processo é igual ao K-médias o que não garante que restrições impostas pelos objectos rotulados sejam totalmente satisfeitas. No algoritmo 4.4, o K-médias semeado é descrito sucintamente.

No algoritmo K-médias restringido, o conjunto de sementes S é utilizado para inicializar os centros dos grupos, da mesma forma que no algoritmo K-médias semeado. No entanto, os rótulos dos objectos sementes nunca são alterados nos passos seguintes do algoritmo, isto é, apenas serão alterados rótulos de objectos $x_i \notin S$. No algoritmo 4.5 apresenta-se o K-médias restringido. Como no algoritmo K-médias restringido o rótulo de cada objecto semente é imutável, este deve ser apenas usado quando o conjunto de sementes não tem ruído, isto é, quando todos os rótulos atribuídos dos objectos sementes se encontram correctamente atribuídos.

Algoritmo 4.4: K-médias semeado

Entrada: $\mathcal{X} = \{x_1, \cdots, x_n\}$ - Conjunto de dados, K - Número de grupos pretendido, $S = \{S_1, \cdots, S_K\}$-Conjunto de sementes.
Saída: P - Agrupamento de dados

```
/*Inicializar centros dos grupos x̄_k;
para k ← 1 até K faça
    x̄_k ← (1/|S_k|) Σ_{x_i ∈ S_k} x_i
fim
repita
    /*Atribuir objectos aos grupos;
    para i ← 1 até n faça
        Atribuir x_i a C_{k*}, em que k* = arg min_k ||x_i − x̄_k||^2;
    fim
    /*Estimar novos centros para os grupos;
    para k ← 1 até K faça
        x̄_k ← (1/|C_k|) Σ_{x_i ∈ C_k} x_i
    fim
até não existir alteração dos grupos ;
Devolver P = {C_1, ··· , C_K};
```

Caso contrário, o algoritmo K-médias semeado é uma melhor opção, já que, torna possível que um objecto semente troque de grupo no decorrer do processo de agrupamento.

4.4 Penalização de Violações de Restrições

Os algoritmos de agrupamento de dados apresentados neste secção admitem soluções em que nem todas as restrições são satisfeitas. A ideia principal deste tipo de algoritmos consiste na definição de uma função-objectivo que, para além de considerar as distâncias entre os objectos e respectivos centros de grupos, penaliza a violação de restrições. De seguida são apresentados três algoritmos de agrupamento de dados representativos desta categoria: o PCK-médias [7], o CVQE [22] e o LCVQE [72].

4.4.1 PCK-médias

Basu *et al.* [7] introduzem um novo conceito ao agrupamento de dados com restrições entre pares de objectos: a cada restrição entre pares de objectos é associado um custo de violação da mesma. Esta abordagem, denominada por PPC (*Pairwise Constrained Clustering*), tem como objectivo não só minimizar as distâncias entre os objectos de dados e os centros dos seus grupos, como também, minimizar o custo de violação das restrições.

O algoritmo de agrupamento de dados PCK-médias considera apenas relações entre pares de objectos, mais precisamente, ligações obrigatórias e ligações proibidas, e respectivos custos

Algoritmo 4.5: K-médias restringido

Entrada: $\mathcal{X} = \{x_1, \cdots, x_n\}$ - Conjunto de dados, K - Número de grupos pretendido, $S = \{S_1, \cdots, S_K\}$-Conjunto de sementes.
Saída: P - Agrupamento de dados

```
/*Inicializar centros dos grupos x̄_k;
para k ← 1 até K faça
    x̄_k ← (1/|S_k|) Σ_{x_i ∈ S_k} x_i
fim
repita
    /*Atribuir objectos aos grupos;
    para i ← 1 até n faça
        se x_i ∈ S então
            para k ← 1 até K faça
                se x_i ∈ S_k então
                    Atribuir x_i a C_k;
                fim
            fim
        senão
            Atribuir x_i a C_{k*}, em que k* = arg min_k ||x_i − x̄_k||^2;
        fim
    fim
    /*Estimar novos centros para os grupos;
    para k ← 1 até K faça
        x̄_k ← (1/|C_k|) Σ_{x_i ∈ C_k} x_i
    fim
até não existir alteração dos grupos ;
Devolver P = {C_1, ··· , C_K};
```

de violação. Considere-se o conjunto de restrições de ligação obrigatória $Rest_{=}$, o conjunto de ligação proibida $Rest_{\neq}$ e os conjuntos de custo de violação dessas restrições $W_{=} = \{w_{=_{ij}}\}$ e $W_{\neq} = \{w_{\neq_{ij}}\}$, respectivamente. Considere-se também o rótulo $l_i \in \{k\}_1^K$ atribuído ao objecto $x_i \in \mathcal{X}$ resultante do agrupamento de um conjunto com n objectos de dados, $\mathcal{X} = \{x_1, \cdots, x_n\}$, em K grupos. O custo de violação de uma restrição de ligação obrigatória $(x_i, x_j) \in Rest_{=}$ é calculado por $w_{=_{ij}} I(l_i \neq l_j)$, em que $I(\cdot)$ devolve 1 se a expressão for verdadeira e 0 caso contrário, isto é, se os dois objectos da ligação obrigatória forem atribuídos a grupos diferentes o custo da violação de restrição é dado pelo custo $w_{=_{ij}}$. Analogamente, o custo de violação de uma restrição de ligação proibida $(x_i, x_j) \in Rest_{\neq}$ é calculado por $w_{\neq_{ij}} I(l_i = l_j)$.

Assim, a função-objectivo a minimizar é dada na equação 4.2, em que o objecto de dados x_i é atribuído ao grupo C_{l_i} cujo centro é representado por $\overline{x}_{l_i}$.

$$\jmath_{PPC} = \frac{1}{2}\sum_{x_i \in \mathcal{X}} ||x_i - \overline{x}_{l_i}||^2 + \sum_{(x_i,x_j)\in Rest_=} w_{=_{ij}} I(l_i \neq l_j) + \sum_{(x_i,x_j)\in Rest_\neq} w_{\neq_{ij}} I(l_i = l_j) \tag{4.2}$$

O algoritmo PCK-médias (*Pairwise Constrained K-means*) [7], optimiza de forma gulosa a função-objectivo $\jmath_{PPC}$ (equação 4.2) através de uma modificação do algoritmo K-médias no passo da atribuição dos objectos de dados aos grupos. Como já referido na secção 4.3.1, uma inicialização apropriada do K-médias, atendendo às restrições impostas pelo utilizador, aumenta o desempenho do agrupamento de dados. No passo de inicialização do PCK-médias, o conjunto de restrições de ligação obrigatória, $Rest_=$, é expandido através de relações de transitividade entre os objectos de $Rest_=$ e os restantes objectos. Note-se que, se o conjunto de restrições de ligação obrigatória original for ruidoso, isto é, existirem restrições erradamente definidas pelo utilizador, esta expansão de $Rest_=$ poderá resultar num decréscimo da qualidade do agrupamento de dados. Considere-se λ o número de componentes do conjunto expandido de restrições de ligação obrigatória $Rest_=$, em que cada componente forma um conjunto de vizinhança (objectos ligados) $V_p \in \{V_1, \cdots, V_\lambda\}$. Para cada par de conjuntos de vizinhança, V_p e $V_{p'}$, em que exista pelo menos uma restrição de ligação proibida, são adicionadas restrições de ligação proibida entre cada par de objectos (x_i, x_j), $x_i \in V_p$ e $x_j \in V_{p'}$ expandindo assim o conjunto de ligações proibidas $Rest_\neq$.

Após este pré-processamento, os λ conjuntos de vizinhança, $\{V_1, \cdots, V_\lambda\}$, são utilizados para definir os centros iniciais de cada grupo. Se $\lambda \geq K$, em que K é o número pretendido de grupos, são seleccionados os K conjuntos de vizinhança com mais objectos e definem-se os grupos iniciais como os centros desses conjuntos de vizinhança. Se $\lambda < K$, inicializam-se λ centros de grupos com os centros de cada um dos λ conjuntos de vizinhança. De seguida, é procurado um objecto de dados x_i que possua restrições de ligação proibida a cada um dos λ conjuntos de vizinhança e, caso o objecto exista, x_i é usado para definir o centro de grupo $\lambda + 1$. Finalmente, caso nem todos os centros dos grupos tenham sido inicializados, estes são definidos aleatoriamente.

O PCK-médias consiste num processo iterativo (algoritmo 4.6) em que, inicialmente, os objectos de dados são atribuídos a um grupo e, posteriormente, os centros de grupo são actualizados. No passo de atribuição dos objectos aos grupos, cada objecto de dados x_i é atribuído ao grupo que minimiza a soma da distância de x_i ao centro do grupo com o custo de violação de restrições imposta por essa atribuição. Repare-se que este passo é dependente da ordem pela qual os objectos são atribuídos aos grupos, já que, os subconjuntos de objectos de $Rest_=$ e $Rest_\neq$ existentes em cada grupo podem variar com a atribuição de um objecto.

Algoritmo 4.6: PCK-médias

Entrada: $\mathcal{X} = \{x_1, \cdots, x_n\}$ - Conjunto de dados, K - Número de grupos pretendido, $Rest_=$ - Restrições de ligação obrigatória, $Rest_{\neq}$ - Restrições de ligação proibida, $W_= = \{w_{=_{ij}}\}_{\forall i,j:(x_i,x_i)\in Rest_=}$- Custos da violação das ligações obrigatórias, $W_{\neq} = \{w_{\neq_{ij}}\}_{\forall i,j:(x_i,x_i)\in Rest_{\neq}}$- Custos da violação das ligações proibidas

Saída: P - Agrupamento de dados

```
/*Inicializar centros dos grupos x̄_k;
Criar os λ vizinhos {V_1, ··· , V_λ} a partir de Rest_= e Rest_≠;
Ordenar os índices p por ordem decrescente da cardinalidade de V_p;
se λ ≥ K então
    para k ← 1 até K faça
        Inicializar x̄_k com o centro de V_k;
    fim
senão
    para k ← 1 até λ faça
        Inicializar x̄_k com o centro de V_k;
    fim
    se ∃ um objecto x_i com restrições de ligação proibida a todos as vizinhanças {V_k}_{k=1}^λ
    então
        Inicializar x̄_k ← x_i;
    fim
    Inicializar os restantes centros de grupo aleatoriamente;
fim
repita
    /*Atribuir objectos aos grupos;
    para i ← 1 até n faça
        Atribuir x_i a C_{k*}, em que k* =
        arg min_k (1/2)||x_i − x̄_k||^2 + Σ_{(x_i,x_j)∈Res_=} w_{=_ij} I(k ≠ l_j) + Σ_{(x_i,x_j)∈Res_≠} w_{≠_ij} I(k = l_j);
    fim
    /*Estimar novos centros para os grupos;
    para k ← 1 até K faça
        x̄_k ← (1/|C_k|) Σ_{x_i∈C_k} x_i
    fim
até não existir alteração dos grupos ;
Devolver P = {C_1, ··· , C_K};
```

4.4.2 CVQE

Nesta secção, apresenta-se mais um algoritmo de agrupamento baseado no algoritmo K-médias e na modificação da sua função-objectivo. A função-objectivo do algoritmo K-médias é equivalente ao *erro de quantificação vectorial* (*Vector Quantization Error* - VQE), pois tenta minimizar iterativamente o *erro de quantificação vectorial*, também denominado por *distorção*. O VQE é definido pelas equações 4.3 e 4.4.

$$VQE = \sum_{j=1}^{K} VQE_j \tag{4.3}$$

$$VQE_j = \frac{1}{2} \sum_{x_i \in C_j} d(\overline{x}_j, x_i)^2 \tag{4.4}$$

Como já anteriormente explicado, o algoritmo K-médias itera entre a atribuição dos objectos ao grupo cujo centro se encontra mais próximo e a actualização dos centros dos grupos. Este último passo pode ser obtido derivando o erro (equação 4.4) relativo ao grupo C_j e igualando-o a 0, tal como mostram as equações 4.5 e 4.6.

$$\frac{\partial VQE_j}{\partial \overline{x}_j} = \sum_{x_i \in C_j} d(\overline{x}_j, x_i)^2 = 0 \tag{4.5}$$

$$\overline{x}_j = \frac{\sum_{x_i \in C_j}(x_i)}{|C_j|} \tag{4.6}$$

Davidson e Ravi [22] criaram uma nova versão do K-médias, denominado *erro de quantificação vectorial restringido* (*Constrained Vector Quantization Error* - CVQE), com o intuito de incorporar restrições dos tipo ligação obrigatória e proibida. A ideia principal do seu algoritmo consiste na criação de uma nova função-objectivo diferenciável que considere estas restrições. A nova função-objectivo é apresentada seguidamente:

$$\begin{aligned} CVQE_j &= \frac{1}{2} \sum_{x_i \in C_j} d(\overline{x}_j, x_i)^2 \\ &+ \frac{1}{2} \sum_{i=1, g_i}^{NumRest} [d(\overline{x}_j, \overline{x}_{g_i'})^2 I(g_i \neq g_i')]^{L_o} \times [d(\overline{x}_j, \overline{x}_{h_{g_i'}})^2 I(g_i = g_i')]^{1-L_o} \end{aligned} \tag{4.7}$$

em que g_i e g_i' correspondem aos índices dos grupos mais próximos do primeiro e segundo objecto na i-ésima ligação obrigatória/proibida, h_i retorna o índice do grupo cujo centro se encontra mais próximo de $\overline{x}_i$, $NumRest = |Rest_=| + |Rest_{\neq}|$ é o número de total de pares de objectos restringidos, $L_o = 1$ se a restrição for uma ligação obrigatória e $L_o = 0$ no caso contrário. A primeira parte da função-objectivo da equação 4.7 corresponde à distorção enquanto

que os restantes termos são os erros associados à violação de ligações obrigatórias ($L_o = 1$) e proibidas($L_o = 0$).

O primeiro passo da nova versão do algoritmo de agrupamento de dados K-médias tem de minimizar a função-objectivo definida na equação 4.7, o que é alcançado atribuindo os objectos aos grupos de forma a minimizar o primeiro termo da mesma equação. Os objectos não restringidos são associados ao grupo mais próximo, analogamente ao algoritmo K-médias. Para os restantes objectos, para cada par de objectos envolvido numa restrição, são testadas todas as possibilidades de atribuição desses objectos a grupos e é escolhida a que menos aumenta o valor da função-objectivo. A regra de atribuição dos objectos a grupos é formalizada na equação 4.8.

$$\begin{aligned} \forall x_a \in \{Rest_= \cup Rest_{\neq}\} &: \arg\min_j d(x_a, \overline{x}_j)^2 \\ \forall (x_a, x_b) \in Rest_= &: \arg\min_{i,j} d(x_a, \overline{x}_i)^2 + d(x_b, \overline{x}_j)^2 + I(x_a \neq x_b) d(\overline{x}_i, \overline{x}_j) \\ \forall (x_a, x_b) \in Rest_{\neq} &: \arg\min_{i,j} d(x_a, \overline{x}_i)^2 + d(x_b, \overline{x}_j)^2 + I(x_a = x_b) d(\overline{x}_i, \overline{x}_{h_j}) \end{aligned} \tag{4.8}$$

O segundo passo do algoritmo de agrupamento CVQE consiste na actualização dos centros dos grupos, com intuito de minimizar o *erro de quantificação vectorial restringido*. Para isso, é necessário calcular o diferencial de primeira ordem relativo ao erro e resolvê-lo quando igualado a 0 (equação 4.9).

$$\begin{aligned} \frac{\partial CVQE_j}{\partial \overline{x}_j} &= \sum_{x_i \in C_j} d(\overline{x}_j - x_i) \\ &+ \sum_{i=1, g_i=j, I(g_i, g_i')=0}^{|Rest_=|} d(\overline{x}_j, \overline{x}_{g_i}) \\ &+ \sum_{i=|Rest_=|+1, g_i=j, I(g_i, g_i')=1}^{|Rest_{\neq}|} d(\overline{x}_j, \overline{x}_{h_{g_i}}) \\ &= 0 \end{aligned} \tag{4.9}$$

Resolvendo para $\overline{x}_j$ obtém-se a seguinte regra de actualização dos centros dos grupos:

$$\overline{x}_j = \frac{\sum_{x_i \in C_j} \left[x_i + \sum_{(x_i, x_a) \in Rest_=, g_i \neq g_a} \overline{x}_{g_a} + \sum_{(x_i, x_a) \in Rest_{\neq}, g_i = g_a} \overline{x}_{h_{g_a}} \right]}{|C_j| + \sum_{x_i \in C_j, (x_i, x_a) \in Rest_=, g_i \neq g_a} 1 + \sum_{x_i \in C_j, (x_i, x_a) \in Rest_{\neq}, g_i = g_a} 1} \tag{4.10}$$

Na actualização dos centros dos grupos, seguindo a regra apresentada na equação 4.10, quando uma ligação obrigatória $(x_i, x_j) \in Rest_=$, com $x_i \in C_l$, é violada, o grupo C_l é movido em direcção ao grupo C_m que contém o objecto $x_j \in C_m$. Identicamente, quando é violada uma ligação proibida $(x_i, x_j) \in Rest_{\neq}, x_i \in C_l, x_j \in C_l$, o centro do grupo de C_l é movido em direcção

ao grupo mais próximo C_m para que, eventualmente, um dos objectos seja atribuído a C_m e a restrição seja satisfeita.

No algoritmo 4.7 encontra-se o pseudo-código do algoritmo de agrupamento de dados CVQE.

Algoritmo 4.7: CVQE

Entrada: $\mathcal{X} = \{x_1, \cdots, x_n\}$ - Conjunto de dados, K - Número de grupos pretendido, $Rest_=$ - Restrições de ligação obrigatória, $Rest_{\neq}$ - Restrições de ligação proibida, $W_= = \{w_{=_{ij}}\}_{\forall i,j:(x_i,x_i)\in Rest_=}$- Custos da violação das ligações obrigatórias, $W_{\neq} = \{w_{\neq_{ij}}\}_{\forall i,j:(x_i,x_i)\in Rest_{\neq}}$- Custos da violação das ligações proibidas

Saída: P - Agrupamento de dados

Inicializar os centros de grupo, $\overline{x}_k$, aleatoriamente;
repita
  `/*Atribuir objectos aos grupos;`
  para $a \leftarrow 1$ **até** n **faça**
    Atribuir x_a a um grupo, tal que,

$$\begin{aligned}
\forall x_a \in \{Rest_= \cup Rest_{\neq}\} &: \arg\min_j d(x_a, \overline{x}_j)^2 \\
\forall (x_a, x_b) \in Rest_= &: \arg\min_{i,j} d(x_a, \overline{x}_i)^2 + d(x_b, \overline{x}_j)^2 + I(l_a \neq l_b) d(\overline{x}_i, \overline{x}_j) \\
\forall (x_a, x_b) \in Rest_{\neq} &: \arg\min_{i,j} d(x_a, \overline{x}_i)^2 + d(x_b, \overline{x}_j)^2 + I(l_a = l_b) d(\overline{x}_i, \overline{x}_{h_j})
\end{aligned}$$

  fim
  `/*Estimar novos centros para os grupos;`
  para $k \leftarrow 1$ **até** K **faça**

$$\overline{x}_k = \frac{\sum_{x_i \in C_k} \left[x_i + \sum_{(x_i,x_a)\in Rest_=, g_i \neq g_a} \overline{x}_{g_a} + \sum_{(x_i,x_a)\in Rest_{\neq}, g_i = g_a} \overline{x}_{h_{g_a}} \right]}{|C_k| + \sum_{x_i \in C_k, (x_i,x_a)\in Rest_=, g_i \neq g_a} 1 + \sum_{x_i \in C_k, (x_i,x_a)\in Rest_{\neq}, g_i = g_a} 1}$$

  fim
até *não existir alteração dos grupos* ;
Devolver $P = \{C_1, \cdots, C_K\}$;

4.4.3 LCVQE

Pelleg e Baras [72] criaram uma variação do algoritmo de agrupamento CVQE que apelidaram de LCVQE (*Linear-time Constrained Vector Quantization Error*). Este algoritmo difere do CVQE nas regras de atribuição dos objectos a grupos e de actualização dos centros dos grupos. No LCVQE, em vez de se calcular todas as k^2 possibilidades de atribuição de um par de objectos (x_i, x_j) a grupos, apenas são verificadas três possíveis atribuições:

1. Atribuição dos objectos x_i e x_j aos grupos mais próximos;

2. Atribuição dos objectos x_i e x_j ao grupo mais próximo de x_i;

3. Atribuição dos objectos x_i e x_j ao grupo mais próximo de x_j.

A penalização de violação das ligações obrigatórias continua a ser a penalização do algoritmo CVQE. No entanto, a penalização de violação das ligações proibidas passa a ser a distância entre o centro do grupo a que os objectos são atribuídos (C_k) e o centro do grupo mais próximo do objecto (x_i ou x_j) mais afastado de C_k. As regra de atribuição dos objectos aos grupos e da actualização dos centros dos grupos são apresentadas nas equações 4.11 e 4.12, respectivamente.

$$\begin{aligned}
\forall x_a \in \{Rest_= \cup Rest_{\neq}\} \quad &: \quad \arg\min_j d(x_a, \overline{x}_j)^2 \\
\forall (x_a, x_b) \in Rest_= \quad &: \quad \arg\min_{[i=g_a, j=g_b], [i=j=g_a], [i=j=g_b]} d(x_a, \overline{x}_i)^2 + d(x_b, \overline{x}_j)^2 \\
&+ \quad I(l_a \neq l_b) d(\overline{x}_i, \overline{x}_j) \\
\forall (x_a, x_b) \in Rest_{\neq} \quad &: \quad \arg\min_{[i=g_a, j=g_b], [d(x_a, \overline{x}_{g_a}) < d(x_b, \overline{x}_{g_b}) : i=j=g_a)]} d(x_a, \overline{x}_i)^2 + d(x_b, \overline{x}_j)^2 \\
&+ \quad I(l_a = l_b) d(\overline{x}_i, \overline{x}_{g_b})
\end{aligned} \tag{4.11}$$

$$\overline{x}_k = \frac{\sum_{x_i \in C_k} \left[x_i + \sum_{(x_i, x_a) \in Rest_=, g_i \neq g_a} \overline{x}_{g_a} + \sum_{(x_i, x_a) \in Rest_{\neq}, g_i = g_a, d(x_i, \overline{x}_{l_i}) < d(x_a, \overline{x}_{l_a})} \overline{x}_{g_a} \right]}{|C_k| + \sum_{x_i \in C_k, (x_i, x_a) \in Rest_=, g_i \neq g_a} 1 + \sum_{x_i \in C_k, (x_i, x_a) \in Rest_{\neq}, g_i = g_a} 1} \tag{4.12}$$

O algoritmo de agrupamento de dados LCVQE é descrito no algoritmo 4.8.

4.5 Edição de Distância

Outra abordagem para a integração de restrições no agrupamento de dados consiste na edição da medida de distância. Os algoritmos de agrupamento desta categoria, para além de tentarem satisfazer as restrições impostas, tentam generalizar essas restrições ao nível do espaço dos atributos de dados. A figura 4.1 exemplifica a ideia na qual se baseia a edição de distância. A figura 4.1 a) representa um conjunto de dados que se pretende agrupar em dois grupos, existindo duas restrições de ligação obrigatórias. Uma possível solução é apresentada na figura 4.1 b). Como se pode verificar, as restrições impostas foram cumpridas na sua totalidade, apesar do agrupamento de dados apresentado não ser muito intuitivo. Por outro lado, a figura 4.1 c) apresenta também uma solução que satisfaz todas as restrições, mas que para além de apenas satisfazer as restrições impostas generaliza essas restrições tendo em conta o seguinte raciocínio: se dois objectos de dados devem ser agrupados conjuntamente por estarem relacionados com uma

Algoritmo 4.8: LCVQE

Entrada: $\mathcal{X} = \{x_1, \cdots, x_n\}$ - Conjunto de dados, K - Número de grupos pretendido, $Rest_=$ - Restrições de ligação obrigatória, $Rest_{\neq}$ - Restrições de ligação proibida, $W_= = \{w_{=_{ij}}\}_{\forall i,j:(x_i,x_i)\in Rest_=}$- Custos da violação das ligações obrigatórias, $W_{\neq} = \{w_{\neq_{ij}}\}_{\forall i,j:(x_i,x_i)\in Rest_{\neq}}$- Custos da violação das ligações proibidas

Saída: P - Agrupamento de dados

Inicializar os centros de grupo,$\overline{x}_k$, aleatoriamente;
repita
  `/*Atribuir objectos aos grupos;`
  para $a \leftarrow 1$ **até** n **faça**
    Atribuir x_a a um grupo, tal que,

$$\begin{aligned}
\forall x_a \in \{Rest_= \cup Rest_{\neq}\} \quad &: \quad \arg\min_j d(x_a, \overline{x}_j)^2 \\
\forall (x_a, x_b) \in Rest_= \quad &: \quad \arg\min_{[i=g_a, j=g_b],[i=j=g_a],[i=j=g_b]} d(x_a, \overline{x}_i)^2 \\
&+ \quad d(x_b, \overline{x}_j)^2 + I(l_a \neq l_b) d(\overline{x}_i, \overline{x}_j) \\
\forall (x_a, x_b) \in Rest_{\neq} \quad &: \quad \arg\min_{[i=g_a, j=g_b],[d(x_a,\overline{x}_{g_a})<d(x_b,\overline{x}_{g_b}):i=j=g_a)]} d(x_a, \overline{x}_i)^2 \\
&+ \quad d(x_b, \overline{x}_j)^2 + I(l_a = l_b) d(\overline{x}_i, \overline{x}_{h_j})
\end{aligned}$$

  fim
  `/*Estimar novos centros para os grupos;`
  para $k \leftarrow 1$ **até** K **faça**

$$\overline{x}_k = \frac{\sum_{x_i \in C_k} \left[x_i + \sum_{(x_i,x_a)\in Rest_=, g_i \neq g_a} \overline{x}_{g_a} + \sum_{(x_i,x_a)\in Rest_{\neq}, g_i = g_a, d(x_i,\overline{x}_{l_i}) < d(x_a,\overline{x}_{l_a})} \right]}{|C_k| + \sum_{x_i \in C_k,(x_i,x_a)\in Rest_=, g_i \neq g_a} 1 + \sum_{x_i \in C_k,(x_i,x_a)\in Rest_{\neq}, g_i = g_a} 1}$$

  fim
até *não existir alteração dos grupos* ;
Devolver $P = \{C_1, \cdots, C_K\}$;

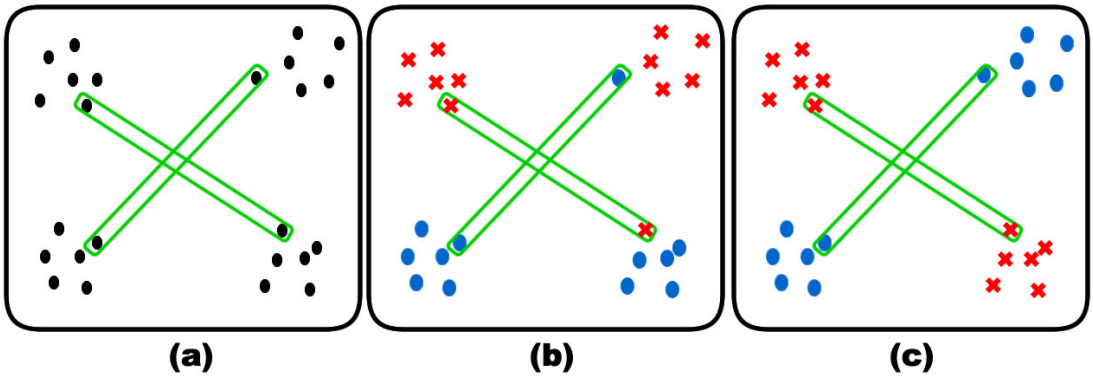

Figura 4.1: Generalização ao nível do espaço de ligações obrigatórias - a) conjunto de dados; b) restrição satisfeita; c) restrição satisfeita e generalizada no espaço.

ligação obrigatória, então os objectos de dados próximos destes também devem ser agrupados no mesmo grupo.

Nas subsecções 4.5.1 e 4.5.2 são apresentados um algoritmo de agrupamento hierárquico e um algoritmo de agrupamento de partição de dados em que o conceito de edição de distância é explorado.

4.5.1 Ligação Completa Restringido

O algoritmo Ligação Completa Restringido [52] (CCL - *Constrained Complete-Link*) extende o bem conhecido algoritmo hierárquico aglomerativo Ligação Completa (*Complete-Link*), permitindo a especificação de ligações obrigatórias e proibidas entre pares de objectos de dados. Dada uma matriz de dissimilaridades, que represente as dissimilaridades entre os pares de objectos do conjunto de dados, e os conjuntos de relações obrigatórias $Rest_{=}$ e proibidas $Rest_{\neq}$, é criada uma nova matriz de dissimilaridades em que as distâncias entre objectos de dados são alteradas, reflectindo as restrições $Rest_{=}$ e $Rest_{\neq}$ e as suas implicações nos objectos do conjunto de dados. Em seguida, é aplicado o algoritmo Ligação Completa para se obter o agrupamento do conjunto de dados.

A ideia do algoritmo de agrupamento de dados CCL consiste na distorção do espaço de similaridade, aproximando os objectos de dados que se sabe que pertencem ao mesmo grupo e afastando objectos de dados que pertencem a grupos diferentes. Na figura 4.2 é apresentado um exemplo do efeito de uma ligação obrigatória na distorção do espaço de similaridade.

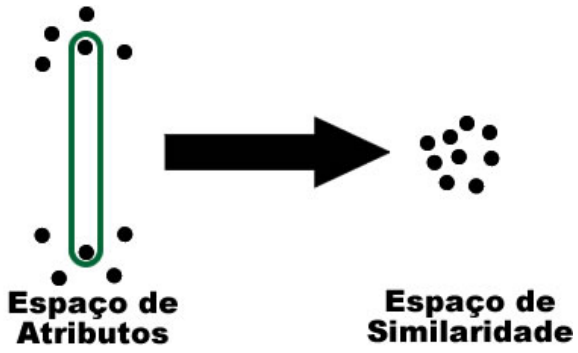

Figura 4.2: Propagação de uma ligação obrigatória - Conjunto de objectos de dados afastados no espaço de atributos são aproximados no espaço de similaridade com a existência de uma ligação obrigatória.

Após esta primeira distorção da matriz de similaridades, o algoritmo de agrupamento CCL propaga as restrições aos objectos de dados não incluídos nos conjuntos $Rest_{=}$ e $Rest_{\neq}$ seguindo as seguintes intuições:

- Se dois objectos de dados x_i e x_j se encontram muito próximos, então os objectos de dados próximos de x_i devem estar próximos de x_j.
- Se um objecto de dados x_i se encontra muito afastado de outro objecto de dados x_j, então os objectos próximos de x_i devem estar afastados de x_j.

A figura 4.3 ilustra as duas intuições supracitadas.

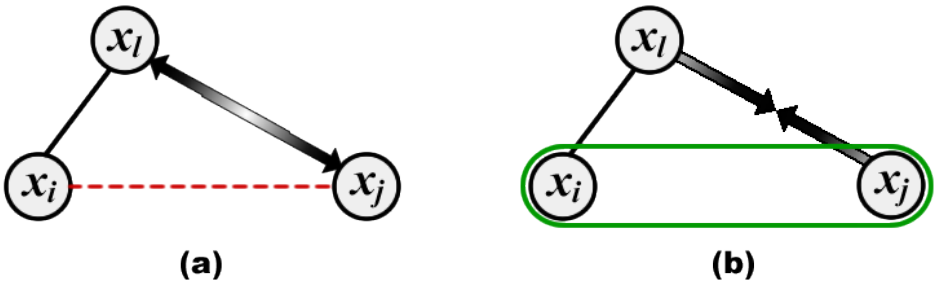

Figura 4.3: Implicações das relações entre pares de objectos - Se x_i e x_l estiverem próximos e se existir uma ligação proibida entre x_i e x_j, x_l é afastado de x_j. Se x_i e x_l estiverem próximos e existir uma ligação obrigatória entre x_i e x_j, x_l é aproximado de x_j.

O pseudo-código do algoritmo de Ligação Completa Restringido é apresentado no algoritmo 4.9. Inicialmente, são impostas as ligações obrigatórias entre pares de objectos na matriz de dissimilaridades $D \in \mathbb{R}^{n \times n}$, atribuindo o valor 0 a cada entrada na matriz entre pares de objectos com ligações obrigatórias, isto é,$\forall (x_i, x_j) \in Rest_{=}, D_{ij}, D_{ji} = 0$. A propagação das ligações obrigatórias ao resto do conjunto de dados é então efectuada através de uma versão modificada do algoritmo de Floyd-Warshall [34], que calcula os caminhos mais próximos entre todos os pares de objectos do conjunto de dados. Em seguida impõem-se as ligações proibidas atribuindo o valor ∞ às entradas na matriz D correspondentes aos objectos de dados com ligações proibidas, ou seja, $\forall (x_i, x_j) \in Rest_{\neq}, D_{ij}, D_{ji} = \infty$. Note-se que não é necessário propagar as restrições de ligação proibida, já que, a propagação será realizada implicitamente na aplicação do algoritmo de Ligação Completa. Finalmente, o algoritmo de Ligação Completa é aplicado a D para que se obter o dendrograma dos objectos de dados. O dendrograma representa a hierarquia da formação dos grupos. Para se obter um agrupamento de dados basta cortar o dendrograma num determinado nível, por exemplo, num número de grupos pretendido.

4.5.2 MPCK-médias

O algoritmo de agrupamento de dados MPCK-médias [6] para além de penalizar violações de ligações obrigatórias e/ou proibidas entre pares de objectos, efectua também aprendizagem da medida de distância, estendendo assim o algoritmo PCK-médias apresentado na subsecção 4.4.1. A evolução deste algoritmo baseia-se no trabalho de Xing *et al.* [93] sobre a parametrização da distância euclidiana.

A medida de distância euclidiana pode ser parametrizada através de uma matriz de ponderações $A \in \mathbb{R}^{d \times d}$ simétrica e positiva da seguinte forma:

$$\|x_i - x_j\|_A = \sqrt{(x_i - \overline{x}_{l_i})^{\mathrm{T}} A (x_j - \overline{x}_{l_j})} \tag{4.13}$$

Algoritmo 4.9: Ligação Completa Restringido

Entrada: $D \in \mathbb{R}^{n \times n}$ - Matriz de dissimilaridades entre n objectos de dados, $Rest_{=}$ - Restrições de ligação obrigatória, $Rest_{\neq}$ - Restrições de ligação proibida

```
/*Realizar a distorção de d(·) de acordo com Rest_= e Rest_≠;
/*Impor ligações obrigatórias;
para cada (x_i, x_j) ∈ Rest_= faça
    D_ij, D_ji = 0;
fim
/*Propagar ligações obrigatórias;
I = i : ∃j ≠ i, (x_i, x_j) ∈ Rest_= ;
para cada l ∈ I faça
    para i = 1 até n faça
        para j = 1 até n faça
            D_ij = min(D_ij, D_il + D_lj);
        fim
    fim
fim
para cada (i, j) tal que D_ij = 0 faça
    Rest_= = Rest_=∪ {x_i, x_j}
fim
/* Impor ligações proibidas (Não é necessário propagar.  Estas restrições
são propagadas implicitamente.);
para cada (x_i, x_j) ∈ Rest_≠ faça
    para cada (x_j, x_k) ∈ Rest_= faça
        D_ik, D_jk = ∞;
    fim
fim
/*Agrupar objectos de dados usando o algoritmo de Ligação Completa;
Grupos = {};
para cada x_i ∈ X faça
    C_i = {x_i};
    Grupos = Grupos ∪ {C_i};
fim
Dendograma={};
/*Inicializar distâncias entre grupos ;
∀(i, j) ∈ {1, ··· , n}, δ(C_i, C_j) = D_ij;
enquanto |Grupos| > 1 faça
    (C_l, C_m) = arg min_{(C_i,C_j)∈Grupos} δ(C_i, C_j);
    Dendograma = Dendograma ∪ {(C_l, C_m)};
    C_novo = {C_l ∪ C_m};
    Grupos = Grupos ∪ C_novo;
    para cada C_i ∈ Grupos faça
        δ(C_i, C_novo) = max{δ(C_i, C_l), δ(C_i, C_m)};
    fim
fim
```

em que x_i é o vector de atributos de um objecto de dados e $\overline{x}_{l_i}$ o vector médio do grupo a que pertence. Se a matriz A for diagonal, então cada atributo de dados é ponderado com o valor da entrada da diagonal respectiva. Caso contrário, são criados implicitamente novos atributos de dados, que são combinações lineares dos atributos originais de dados.

Usar o algoritmo de agrupamento de dados K-médias de forma a minimizar a distância, definida na equação 4.13, para todos os objectos de dados é equivalente a minimizar a seguinte função-objectivo:

$$\jmath_M = \sum_{x_i \in \mathcal{X}} ||x_i - \overline{x}_{l_i}||_A^2 - \log(\det(A)) \tag{4.14}$$

em que $||x_i - \overline{x}_{l_i}||_A^2$ é a distância quadrática entre x_i e o centro do grupo a que foi atribuído x_{l_i}, segundo a medida de distância parametrizada pela matriz A.

A combinação das equações 4.2 e 4.14 originam a seguinte função-objectivo que minimiza simultaneamente a dispersão dos grupos segundo a medida de distância aprendida e a violação de restrições entre pares de objectos de dados:

$$\begin{aligned} \jmath_{combinada} = \sum_{x_i \in \mathcal{X}} ||x_i - \overline{x}_{l_i}||_A^2 + \sum_{(x_i,x_j) \in Rest_=} w_{=_{ij}} I(l_i \neq l_j) \\ + \sum_{(x_i,x_j) \in Rest_\neq} w_{\neq_{ij}} I(l_i = l_j) - \log(\det(A)) \end{aligned} \tag{4.15}$$

Geralmente, as ponderações $W_=$ e $W_\neq$ são uniformes, pelo que a violação das restrições são sempre tratadas de igual forma. No entanto, o custo de penalização de uma ligação obrigatória deve ser mais elevado se os dois objectos de dados se encontrarem próximos do que se os dois objectos se encontrarem afastados. O custo de penalização das ligação proibidas devem também ser elevadas se a distância entre os objectos de dados for elevada e reduzidas se essa distância for escassa. Para considerar esta intuição, as ponderações $W_=$ e $W_\neq$ são multiplicadas por funções de penalização, $f_=(x_i, x_j)$ e $f_\neq(x_i, x_j)$ respectivamente, definidas por:

$$f_=(x_i, x_j) = \max(\alpha_{min}, \alpha_{max} - ||x_i - x_j||_A^2) \tag{4.16}$$

$$f_\neq(x_i, x_j) = \min(\alpha_{min} + ||x_i - x_j||_A^2, \alpha_{max}) \tag{4.17}$$

em que α_{min} e α_{max} são constantes não negativas que correspondem, respectivamente, aos valores mínimo e máximo que a penalização pode tomar.

Assim, a função-objectivo a optimizar é definida pela seguinte equação:

$$\begin{aligned} \jmath_{MPC} = \sum_{x_i \in \mathcal{X}} ||x_i - \overline{x}_{l_i}||_A^2 + \sum_{(x_i,x_j) \in Rest_=} w_{=_{ij}} f_=(x_i, x_j) I(l_i \neq l_j) \\ + \sum_{(x_i,x_j) \in Rest_\neq} w_{\neq_{ij}} f_\neq(x_i, x_j) I(l_i = l_j) - \log(\det(A)) \end{aligned} \tag{4.18}$$

em que $I(\cdot)$ toma valor 1 se a expressão for verdadeira e 0 no caso contrário.

Algoritmo 4.10: MPCK-médias

Entrada: $\mathcal{X} = \{x_1, \cdots, x_n\}$ - Conjunto de dados, K - Número de grupos pretendido, $Rest_=$ - Restrições de ligação obrigatória, $Rest_{\neq}$ - Restrições de ligação proibida, $W_= = \{w_{=_{ij}}\}_{\forall i,j:(x_i,x_i)\in Rest_=\}}$- Custos da violação das ligações obrigatórias, $W_{\neq} = \{w_{\neq_{ij}}\}_{\forall i,j:(x_i,x_i)\in Rest_{\neq}\}}$- Custos da violação das ligações proibidas

Saída: P - Agrupamento de dados

```
/*Inicializar centros dos grupos x̄_k;
Criar os λ vizinhos {V_1, ··· , V_λ} a partir de Rest_= e Rest_≠;
Ordenar os índices p por ordem decrescente da cardinalidade de V_p;
se λ ≥ K então
    para k ← 1 até K faça
        Inicializar x̄_k com o centro de V_k;
    fim
senão
    para k ← 1 até λ faça
        Inicializar x̄_k com o centro de V_k;
    fim
    se ∃ um objecto x_i com restrições de ligação proibida a todos as vizinhanças {V_k}_{k=1}^λ
    então
        Inicializar x̄_k ← x_i;
    fim
    Inicializar os restantes centros de grupo aleatoriamente;
fim
repita
    /*Atribuir objectos aos grupos;
    para i ← 1 até n faça
        Atribuir x_i a C_{k*}, em que k* = arg min_k (1/2)||x_i − x̄_k||²_A − log det(A) +
        Σ_{(x_i,x_j)∈Res_=} w_{=_ij} f_=(x_i, x_j) I(k ≠ l_j) + Σ_{(x_i,x_j)∈Res_≠} w_{≠_ij} f_≠(x_i, x_j) I(k = l_j);
    fim
    /*Estimar novos centros para os grupos;
    para k ← 1 até K faça
        x̄_k ← (1/|C_k|) Σ_{x_i∈C_k} x_i
    fim
    /*Actualizar parametrização da medida de distância;
    A^{-1} = Σ_{x_i∈X} (x_i − x̄_{l_i})(x_i − x̄_{l_i})^T − Σ_{(x_i,x_j)∈Rest_=} w_{=_ij} (x_i − x_j)(x_i − x_j)^T I(l_i ≠ l_j) +
    Σ_{(x_i,x_j)∈Rest_≠} w_{≠_ij} (x_i − x_j)(x_i − x_j)^T I(l_i = l_j)
até não existir alteração dos grupos ;
Devolver P = {C_1, ··· , C_K};
```

O algoritmo de agrupamento de dados com restrições MPCK-médias optimiza de forma gulosa a função-objectivo definida na equação 4.18 através de um algoritmo semelhante ao K-médias, tal como apresentado no algoritmo 4.10. A inicialização dos centros dos grupos é realizada tal como definido no algoritmo PCK-médias, apresentado na secção 4.4.1. O MPCK-médias itera

entre a atribuição dos objectos de dados a grupos, a estimação dos novos centros dos grupos e a aprendizagem da parametrização da medida de distância.

No passo de atribuição dos objectos a grupos, cada objecto x_i é atribuído a um grupo de forma a minimizar a soma das distâncias de x_i ao respectivo centro e do custo das violações de restrições. No passo de estimação dos novos centros de grupos $\overline{x}_k$, as restrições $Rest_=$ e $Rest_{\neq}$ não são consideradas, sendo cada centro de grupo actualizado com o vector médio dos objectos que lhe foram atribuídos:

$$\overline{x}_k \leftarrow \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i \tag{4.19}$$

Finalmente, no passo de aprendizagem da medida de distância, a matriz A é actualizada com o intuito de minimizar a função-objectivo $\jmath_{MPC}$ (equação 4.18). A actualização da matriz A é calculada com a derivada parcial $\frac{\partial \jmath_{MPC}}{\partial A}$ igualada a 0, obtendo-se:

$$\begin{aligned} A = \quad (& \sum_{x_i \in \mathcal{X}} (x_i - \overline{x}_{l_i})(x_i - \overline{x}_{l_i})^{\mathrm{T}} \\ - & \sum_{(x_i, x_j) \in Res^*_=} w_{=_{ij}} (x_i - x_j)(x_i - x_j)^{\mathrm{T}} I(l_i \neq l_j) \\ + & \sum_{(x_i, x_j) \in Res^*_{\neq}} w_{\neq_{ij}} (x_i - x_j)(x_i - x_j)^{\mathrm{T}} I(l_i = l_j))^{-1} \end{aligned} \tag{4.20}$$

em que $Res^*_=$ e $Res^*_{\neq}$ são subconjuntos de $Res_=$ e $Res_{\neq}$ que excluem os pares de objectos cujas funções de penalização $f_=$ e $f_{\neq}$ tomam os valores α_{min} e α_{max} respectivamente.

Os autores deste método indicam que é difícil realizar a aprendizagem da matriz completa A, pelo que limitam a aprendizagem à diagonal de A, o que é equivalente a aprender a medida de distância através da ponderação dos atributos. Assim, o m-ésimo elemento da diagonal de A, a_{mm}, corresponde à ponderação do m-ésimo atributo de dados e é actualizado da seguinte forma:

$$\begin{aligned} a_{mm} = \quad (& \sum_{x_i \in \mathcal{X}} (x_{im} - \overline{x}_{l_i})(x_{im} - \overline{x}_{l_i m})^{\mathrm{T}} \\ - & \sum_{(x_{im}, x_{jm}) \in Res^*_=} w_{=_{ij}} (x_{im} - x_{jm})(x_{im} - x_{jm})^{\mathrm{T}} I(l_i \neq l_j) \\ + & \sum_{(x_{im}, x_{jm}) \in Res^*_{\neq}} w_{\neq_{ij}} (x_{im} - x_{jm})(x_{im} - x_{jm})^{\mathrm{T}} I(l_i = l_j))^{-1} \end{aligned} \tag{4.21}$$

O algoritmo MPCK-médias foi posteriormente redefinido [11], passando a ser local a aprendizagem da medida de distância, isto é, em vez de se aprender a matriz A de parametrização de distância geral para o conjunto de dados, é aprendida uma matriz para cada um dos grupos.

4.6 Modificação do Processo de Geração

Os algoritmos de agrupamento apresentados nesta secção assumem que os dados são gerados segundo um modelo probabilístico, sendo o objectivo dos algoritmos estimar os parâmetros desse modelo, considerando tanto os atributos de dados como as restrições existentes.

4.6.1 Agrupamento Probabilístico de Dados com Penalização

Lu e Leen propuseram um modelo de mistura Gaussiana para realizar o agrupamento de dados, em que as preferências do utilizador são incorporadas na forma de restrições difusas entre pares de objectos [57]. As preferências são representadas através da probabilidade Bayesiana de pares de objectos deverem, ou não, ser atribuídos ao mesmo grupo.

O modelo de mistura Gaussiana é definido por

$$Pr(x|\Theta) = \sum_{k=1}^{K} \pi_k Pr(x|k, \theta_k) \tag{4.22}$$

em que $\Theta = (\pi_1, \cdots, \pi_K)$ consiste no conjunto de parâmetros e K corresponde ao número de componentes da mistura. Ao conjunto de dados $\mathcal{X} = \{x_1, \cdots, x_n\}$ é associada uma variável latente $Z = z(x_i), \{i = 1, \cdots, n\}$, correspondente à atribuição dos objectos de dados a grupos. Deste modo obtêm-se os dados completos $(\mathcal{X}, Z)$. A verosimilhança dos dados completos é calculada por

$$Pr(\mathcal{X}, Z|\Theta) = Pr(\mathcal{X}|Z, \Theta)Pr(Z|\Theta). \tag{4.23}$$

As restrições no agrupamento são incorporadas através da manipulação da probabilidade *a priori* $Pr(Z|\Theta)$, que no modelo de mistura Gaussiana tradicional é imediato, já que, esta é calculada por $Pr(Z|\Theta) = \prod_i \pi_{z_i}$. No agrupamento probabilístico com penalização, a probabilidade *a priori* $Pr(Z|\Theta)$ é alterada através de uma função de ponderação $g(Z)$ que toma valores elevados quando a atribuição dos objectos de dados aos grupos Z são de acordo com as restrições, tomando valores baixos quando tal não acontece:

$$Pr(\mathcal{X}, Z|\Theta, G) = \frac{\prod_i \pi_{z_i} g(Z)}{\sum_Z \prod_j \pi_{z_j} g(Z)} \equiv \frac{1}{K} \prod_i \pi_{z_i} g(Z) \tag{4.24}$$

Dada uma determinada atribuição dos objectos aos grupos, a verosimilhança dos dados completos é calculada pela equação (4.25).

$$Pr(\mathcal{X}, Z|\Theta, G) = Pr(\mathcal{X}|Z, \Theta)\frac{1}{K} \prod_i \pi_{z_i} g(Z) = \frac{1}{K} Pr(\mathcal{X}|Z, \Theta) g(Z) \tag{4.25}$$

A verosimilhança dos dados é então a soma das verosimilhanças dos dados completos para todos

os valores que Z pode tomar, ou seja

$$L(\mathcal{X}|\Theta) = Pr(\mathcal{X}, \Theta, G) = \sum_{Z} Pr(\mathcal{X}, Z|\Theta, G) \tag{4.26}$$

que pode ser maximizado através do algoritmo EM [24].

A função de ponderação das atribuições dos objectos aos grupos é definida por

$$g(Z) = \prod_{i,j} \exp(w_{ij}\delta(z_i, z_j)) \tag{4.27}$$

em que w_{ij} é a ponderação associada ao par de objectos (x_i, x_j) e δ é a função delta de Kronecker em que $\delta(z_i, z_j) = 1$ se $z_i = z_j$, caso contrário $\delta(z_i, z_j) = 0$. A ponderação $w_{ij} \in [-\infty, \infty]$, $w_{ij} = w_{ji}$ indica a confiança da atribuição de dois objectos, x_i e x_j, ao mesmo grupo. No caso de restrições de ligação obrigatória, w_{ij} toma um valor positivo enquanto que w_{ij} é negativo se a a restrição for de ligação proibida. Quando não existe nenhuma restrição entre x_i e x_j o valor da ponderação é $w_{ij} = 0$. A confiança depositada numa determinada restrição é quantificada por $|w_{ij}|$, ou seja, quanto mais afastada de 0 a ponderação w_{ij} for, mais confiança é dada à restrição que envolve x_i e x_j.

Para se determinar os parâmetros do modelo Θ^*, que maximizam a verosimilhança $L(\cdot)$, é utilizado o algoritmo EM.

$$\Theta^* = \arg\max_{\Theta} L(\mathcal{X}|\Theta, G) \tag{4.28}$$

O algoritmo EM maximiza iterativamente a função de verosimilhança até atingir um óptimo (geralmente local) usando dois passos: Expectação e Maximização. No passo de expectação a função de verosimilhança é calculada usando os valores obtidos no passo de maximização na iteração anterior $(t-1)$

$$Q(\Theta, \Theta^{(t-1)}) = E_{z|x}(\log(Pr(\mathcal{X}, Y|\Theta, G)|X, \Theta^{(t-1)}), G) \tag{4.29}$$

No passo de maximização os parâmetros são actualizados de forma a maximizar a função de verosimilhança encontrada no passo de expectação. O centros e matriz de covariância óptimos de cada componente do modelo são calculados por

$$\overline{x}_k = \frac{\sum_{i=1}^{n} x_i Pr(k|x_i, \Theta^{(t-1)}, G)}{Pr(k|x_i, \Theta^{(t-1)}, G)} \tag{4.30}$$

$$\Sigma_k = \frac{Pr(k|x_i, \Theta^{(t-1)}, G)(x_i - \overline{x}_k)(x_i - \overline{x}_k)^{\mathrm{T}}}{Pr(k|x_i, \Theta^{(t-1)}, G)} \tag{4.31}$$

Para a actualização das probabilidades *a priori* de cada um dos componentes é necessário en-

contrar

$$\pi \equiv \{\pi_1, \cdots, \pi_K\} = \arg\max_{\pi} \sum_{k=1}^{K} \sum_{i=1}^{n} \log \pi_k Pr(k|x_i, \Theta^{(t-1)}, G) - \log K(\pi) \tag{4.32}$$

No passo de maximização é necessário calcular a probabilidade de pertença aos grupos *a posteriori*, o que no modelo de mistura Gaussiana padrão é trivial mas neste caso torna-se complicado, já que, as distribuições conjuntas *a posteriori* para as relações entre pares de objectos têm de ser calculadas tendo em conta a transitividade das restrições de ligação obrigatória e ligação proibida.

$$Pr(z_i, z_j|\mathcal{X}, \Theta, W) \neq Pr(z_i|\mathcal{X}, \Theta, W) Pr(z_j|\mathcal{X}, \Theta, W) \tag{4.33}$$

Os conjuntos mais pequenos de objectos para os quais as probabilidades de atribuição *a posteriori* podem ser calculados independentemente serão designados por *blocos*. A probabilidade *a posteriori* de um objecto dedados x_i num bloco T é calculada marginalizando a probabilidade *a posteriori* na totalidade do bloco

$$Pr(z_i = k|\mathcal{X}, \Theta, W) = \sum_{Z_T|z_i=k} Pr(Z_T|\mathcal{X}_T, \Theta, W) \tag{4.34}$$

em que a probabilidade *a posteriori* no bloco T dado por

$$Pr(Z_T|\mathcal{X}_T, \Theta, W) = \frac{Pr(Z_T, \mathcal{X}_T|\Theta, W)}{Pr(\mathcal{X}_T|\mathcal{X}_T|\Theta, W)} = \frac{Pr(Z_T, \mathcal{X}_T|\Theta, W)}{\sum_{Z'_T} P(Z'_T, \mathcal{X}_T|\Theta, W)} \tag{4.35}$$

O cálculo da probabilidade *a posteriori* de um objecto num bloco T pode ser bastante custoso pelo que apenas é razoável calcular blocos de tamanho reduzido. Por este motivo, quando o tamanho de um bloco T é elevado uma solução é decompo-lo em blocos mais pequenos e computacionalmente tratáveis. Outra forma para inferir probabilidade *a posteriori* utiliza a amostragem de Gibbs [64] para o seu cálculo. Para mais detalhes da utilização da amostragem de Gibbs neste abordagem, o leitor interessado deverá consultar [57].

4.6.2 HMRF *K*-médias

Basu *et al.* propuseram [8] um algoritmo de agrupamento de dados em que são consideradas, simultaneamente, relações de ligação obrigatória e proibida entre pares de objectos de dados e a aprendizagem de uma medida de distância. O algoritmo proposto, denominado HMRF K-médias, tem como objectivo a minimização de uma função-objectivo derivada do modelo HMRF (*Hidden Markov Random Fields*). Esta abordagem aumenta o desempenho do agrupamento de dados não supervisionado em três aspectos, generalizando o algoritmo MPCK-médias apresentado na subsecção 4.5.2:

- Inicialização melhorada - os centros iniciais dos grupos são obtidos com base em conjuntos de vizinhança de objectos induzidos pelas restrições;

- Atribuição dos objectos de dados aos grupos atendendo a restrições - os objectos são atribuídos aos grupos atendendo não só à minimização de uma medida de distorção, mas também minimizando o número de violações de restrições;

- Aprendizagem iterativa da medida de distância - a medida de distorção é actualizada durante o processo de agrupamento de dados, modificando o espaço para que as restrições sejam satisfeitas.

Nesta abordagem apenas são consideradas restrições entre pares de objectos, nomeadamente, relações de ligação obrigatória e de ligação proibida. Para que estas sejam incorporadas no agrupamento de dados juntamente com a medida de distorção é utilizado o modelo probabilístico HMRF, que é composto pelos seguintes componentes:

- Um campo *escondido* $L = \{l_1, \cdots, l_n\}$ de variáveis aleatórias cujos valores não são observáveis. Este conjunto de variáveis escondidas correspondem aos rótulos de cada objecto de dados que se pretendem obter.

- Um conjunto de variáveis aleatórias observáveis, $\mathcal{X} = \{x_1, \cdots, x_n\}$, que correspondem ao conjunto de n objectos de dados que se pretende agrupar. Pressupõe-se que cada variável aleatória x_i é gerada segundo uma distribuição de probabilidade condicional, $p(x_i|l_i)$ que é determinada pela variável escondida correspondente, l_i.

A cada variável aleatória l_i é associado um conjunto de vizinhos V_i composto por todos os objectos de dados que têm relações, de ligação obrigatória ou de ligação proibida, com o objecto x_i. O campo aleatório definido sobre as variáveis escondidas é um campo aleatório de Markov e a distribuição de probabilidade das variáveis escondidas obedecem a seguinte propriedade de Markov:

$$\forall i, Pr(l_i|L - \{l_i\}) = Pr(l_i|\{l_j : l_j \in V_i\}) \tag{4.36}$$

Assim, a distribuição de probabilidade do valor de l_i para o objecto de dados x_i depende apenas dos rótulos dos objectos que possuem ligações (obrigatórias ou proibidas) com x_i. Considerando que $L = \{l_1, \cdots, l_n\}$ é um evento conjunto, a probabilidade de uma configuração de rótulos pode ser definida por

$$Pr(L) = \frac{1}{Z_1} \exp(-U(L)) = \frac{1}{Z_1} \exp(-\sum_{V_i \in V} U_{V_i}(L)) \tag{4.37}$$

em que V é o conjunto de todas as vizinhanças, Z_1 é uma constante de normalização e $U(L)$ é a função potencial da configuração dos rótulos, que pode ser decomposta nas funções $U_{V_i}(L)$, representantes do potencial de cada vizinhança V_i na configuração dos rótulos L. Como nesta

abordagem apenas existem restrições entre pares de objectos, a probabilidade *a priori* de uma configuração de rótulos define-se por

$$Pr(L) = \frac{1}{Z_1} \exp(-\sum_i \sum_j U(i,j))) \tag{4.38}$$

em que

$$U(i,j) = \begin{cases} f_=(x_i, x_j) & \text{se}(x_i, x_j) \in Rest_= \\ f_{\neq}(x_i, x_j) & \text{se}(x_i, x_j) \in Rest_{\neq} \\ 0 & \text{caso contrário} \end{cases}. \tag{4.39}$$

As funções $f_=(x_i, x_j)$ e $f_{\neq}(x_i, x_j)$ devolvem valores não negativos de forma a penalizar a violação de ligações obrigatórias e ligações proibidas, respectivamente.

Para uma determinada configuração dos rótulos dos objectos, ou seja, das variáveis escondidas, os objectos de dados (correspondentes às variáveis observáveis do modelo HMRF) são gerados segundo distribuições de probabilidade condicional

$$Pr(\mathcal{X}|L) = p(\mathcal{X}, \{\overline{x}_k\}_{k=1}^K) \tag{4.40}$$

em que $p(\mathcal{X}, \{\overline{x}_k\}_{k=1}^K)$ é uma função de densidade de probabilidade parametrizada pelos centros dos grupos $\{\overline{x}_1, \cdots, \overline{x}_K\}$. A probabilidade *a posteriori* total de uma configuração de rótulos L é $Pr(L|\mathcal{X}) \propto Pr(L)Pr(\mathcal{X}|L)$ sendo $Pr(\mathcal{X})$ uma constante C. Desta forma, encontrar a configuração Máxima *A Posteriori* (MAP) do modelo HMRF é equivalente a maximizar a probabilidade *a posteriori*

$$Pr(L|\mathcal{X}) = (\frac{1}{Z_2} \exp(-\sum_i \sum_j U(i,j))) \cdot p(\mathcal{X}, \{\overline{x}_k\}_{k=1}^K) \tag{4.41}$$

em que $Z_2 = CZ_1$. A probabilidade *a posteriori* $Pr(L|\mathcal{X})$ (equação 4.41) tem dois componentes. O primeiro verifica a satisfação das restrições impostas pelo utilizador através da avaliação das atribuições dos objectos ao grupo. O segundo componente estima a probabilidade da geração dos objectos de dados usando as distribuições condicionais, que são parametrizados pelos centros dos grupos e pela medida de distorção

$$p(\mathcal{X}, \{\overline{x}_k\}_{k=1}^K) = \frac{1}{Z_3} \exp(-\sum_{x_i \in \mathcal{X}} d(x_i, \overline{x}_{l_i})) \tag{4.42}$$

em que $d(x_i, \overline{x}_{l_i})$ é uma medida de distorção entre x_i e $\overline{x}_{l_i}$) e Z_3 é uma constante de normalização.

Como é pretendido que as restrições impostas pelo utilizador influenciem a medida de distorção, a penalização de violação de uma restrição de ligação obrigatória, $f_=(x_i, x_j)$ deve ser maior quando os dois objectos de dados x_i e x_j se encontram distantes e menor quando se encontram próximos. Da mesma forma, a penalização para uma restrição de ligação proibida,

$f_{\neq}(x_i, x_j)$ deve ser maior quando x_i e x_j se encontram próximos e menor quando se encontram distantes. Assim, as funções de penalização devem estar na forma

$$f_{=}(x_i, x_j) = w_{=_{ij}}\Phi_d(x_i, x_j)I(l_i \neq l_j), f_{\neq}(x_i, x_j) = w_{\neq_{ij}}(\Phi_{d_{max}} - \Phi_d(x_i, x_j))I(l_i = l_j) \quad (4.43)$$

em que Φ_d é uma função escalar de penalização monotonamente crescente de acordo com a função de distorção, $\Phi_{d_{\max}}$ é o valor máximo de Φ_d para o conjunto de dados e $w_{=_{ij}}$ e $w_{\neq_{ij}}$ são os custos de violação de restrições que envolvam os objectos x_i e x_j. Tendo em conta a equação 4.41, a função-objectivo para o agrupamento de dados pode ser definida por

$$\begin{aligned} \jmath_{HMRF} = &\sum_{x_i \in \mathcal{X}} d(x_i, \overline{x}_{l_i}) + \sum_{(x_i, x_j) \in Rest_{=}} w_{=_{ij}}\Phi_d(x_i, x_j)I(l_i \neq l_j) \\ &+ \sum_{(x_i, x_j) \in Rest_{\neq}} w_{\neq_{ij}}(\Phi_{d_{\max}} - \Phi_d)(x_i, x_j)I(l_i = l_j) + \log Z \end{aligned} \quad (4.44)$$

em que $Z = Z_2 Z_3$. Desta forma, o objectivo do algoritmo de agrupamento de dados com restrições é minimizar $\jmath_{HMRF}$ estimando os parâmetros $\{\overline{x}_1, \cdots, \overline{x}_K\}$, L e $d(\cdot)$. Para isso, o algoritmo K-médias é adaptado originando o HMRF K-médias que se encontra descrito no algoritmo 4.11.

É sabido que a uma boa inicialização dos centros dos grupos é fundamental para o bom desempenho dos algoritmos de agrupamento de partição, como é o caso do K-médias [5]. Nesta abordagem, a inicialização dos centros dos grupos depende não só das restrições do utilizador como também dos objectos de dados que não se encontram restringidos de forma alguma. O processo de inicialização desenrola-se em dois passos: inferência da vizinhança e selecção dos grupos. O primeiro passo é em tudo semelhante ao já descrito na subsecção 4.4.1, em que se encontram λ conjuntos de vizinhança $\{V_1, \cdots, V_\lambda\}$, pelo que aqui apenas se descreverá o passo de selecção dos grupos.

Os λ conjuntos de vizinhança, $\{V_1, \cdots, V_\lambda\}$, obtidos no primeiro passo, são utilizados para definir os centros dos grupos iniciais. Se $\lambda = K$, os K centros dos grupos são os centros de cada vizinhança V_i, equanto que se $\lambda < K$, λ centros de grupos são inicializados com os centros das λ vizinhanças $\{V_1, \cdots, V_\lambda\}$, sendo os restantes centros obtidos aleatoriamente. Por sua vez, se $\lambda > K$, K centros de grupos são seleccionados usando a medida de distorção $d(\cdot)$ utilizada para o agrupamento de dados. O objectivo é encontrar K centros de grupos separados o mais possível segundo $d(\cdot)$. Cada vizinhança V_i tem uma ponderação associada, proporcional ao número de objectos que engloba. A distância entre dois centros de vizinhanças é calculada com o valor da medida de distorção entre os dois centros de vizinhança multiplicado pelas ponderações de cada

Algoritmo 4.11: HMRF K-médias

Entrada: $\mathcal{X} = \{x_1, \cdots, x_n\}$ - Conjunto de dados, K - Número de grupos pretendido, $Rest_{=}$ - Restrições de ligação obrigatória, $Rest_{\neq}$ - Restrições de ligação proibida, $W_{=} = \{w_{=_{ij}}\}_{\forall i,j:(x_i,x_i)\in Rest_{=}}$- Custos da violação das ligações obrigatórias, $W_{\neq} = \{w_{\neq_{ij}}\}_{\forall i,j:(x_i,x_i)\in Rest_{\neq}}$- Custos da violação das ligações proibidas, $d(\cdot)$ - Medida de distorção parametrizável

Saída: P - Agrupamento de dados

```
/*Inicializar centros dos grupos x̄_k;
Criar os λ vizinhos {V_1, ··· , V_λ} a partir de Rest_= e Rest_≠;
Ordenar os índices p por ordem decrescente da cardinalidade de V_p;
se λ ≥ K então
    para k ← 1 até K faça
        Inicializar x̄_k com o centro de V_k;
    fim
senão
    para k ← 1 até λ faça
        Inicializar x̄_k com o centro de V_k;
    fim
    se ∃ um objecto x_i com restrições de ligação proibida a todos as vizinhanças {V_k}_{k=1}^{λ}
    então
        Inicializar x̄_k ← x_i;
    fim
    Inicializar os restantes centros de grupo aleatoriamente;
fim
/*Optimizar função-objectivo J_HMRF*/;
Definir t ← 0;
repita
    Expectação - Dado o conjunto dos centros dos grupos, {x̄_1^t, ··· , x̄_K^t}, actualizar os
    rótulos dos objectos {l_1^{t+1}, ··· , l_n^{t+1}} minimizando J_HMRF;
    Maximização (A) - Dado o conjunto de rótulos {l_1^{t+1}, ··· , l_n^{t+1}}, actualizar os
    centros dos grupos {x̄_1^{t+1}, ··· , x̄_K^{t+1}} e respectivo agrupamento de dados P
    minimizando J_HMRF;
    Maximização (B) - Actualizar a medida de distorção d(·) para reduzir J_HMRF;
    Definir t ← t + 1;
até não existir alteração dos grupos ;
Devolver P;
```

vizinhança. Assim, os centros seleccionados serão relativamente afastados e as suas vizinhanças serão compostas por um grande número de objectos de dados. O processo de selecção dos centros funciona da seguinte forma: inicialmente, o centro cuja vizinhança tem mais objectos é seleccionado e marcado como já visitado; em seguida, calcula-se qual o centro ainda não visitado que se encontra mais afastado dos centros já seleccionado segundo a medida de distorção $d(\cdot)$ ponderada pelas suas vizinhanças; caso surja um empate, este é resolvido escolhendo o grupo mais afastado do centro do conjunto de dados; este processo continua até terem sido encontrados K centros de vizinhanças, sendo esses centros usados para inicializar o algoritmo de agrupamento.

No passo de expectação, os objectos de dados são atribuídos aos grupos seguindo as estimativas actuais dos centros dos grupos. Como no modelo HMRF existe interacção entre os rótulos dos objectos, definida pelo campo aleatório das variáveis escondidas, o cálculo da atribuição óptima é computacionalmente intratável, pelo que é usado o algoritmo guloso ICM (*Iterated Conditional Modes*) para se obter uma boa aproximação. Este algoritmo atribui os objecto de dados a grupos seguindo uma ordem aleatória. Cada objecto de dados x_i é atribuído ao grupo cujo centro $\overline{x}_k$ minimiza a sua contribuição para a função-objectivo $\jmath_{HMRF}(x_i, \overline{x}_k)$

$$\begin{aligned}\jmath_{HMRF}(x_i, \overline{x}_k) = d(x_i, \overline{x}_k) &+ \sum_{(x_i,x_j)\in Rest_=} w_{=ij}\Phi_d(x_i, x_j)I(k \neq l_i) \\ &+ \sum_{(x_i,x_j)\in Rest_\neq} w_{\neq ij}\Phi_d(x_i, x_j)I(k = l_i)\end{aligned} \tag{4.45}$$

O passo de maximização é constituído por duas etapas: primeiro, os centros dos grupos, $\{\overline{x}_1, \cdots, \overline{x}_K\}$, são actualizados para optimizar a função-objectivo $\jmath_{HMRF}$; em seguida, actualizam-se os parâmetros da medida de distorção. Estas actualizações são dependentes das medidas de distorção utilizadas, que não são abordadas neste documento. O leitor interessado poderá consultar [8] onde são apresentadas duas medidas de distorção parametrizáveis, mais precisamente, a similaridade medida pelo co-seno parametrizável e a I-Divergência parametrizável.

4.7 Sumário

Neste capítulo, foram apresentados alguns dos principais algoritmos de agrupamento de dados com a capacidade de incorporar restrições ao nível dos objectos. Os algoritmos de agrupamento de dados descritos são baseados em varias ideias e intuições, tais como, a simples impossibilidade de violação de restrições, a utilização de um subconjunto de objectos rotulados para inicializar os centros de algoritmos de agrupamento de partição, a modificação da função-objectivo de forma a penalizar a violação de restrições, a aprendizagem da medida de distância e, finalmente, a modificação do processo de geração de dados em modelos probabilísticos.

No capítulo 6, *Avaliação de Algoritmos de Agrupamento de Dados e Métodos de Combinação*, é realizada a avaliação de vários dos algoritmos de agrupamento apresentados neste capítulo,

com o objectivo de comparar os seus desempenhos, em conjuntos de dados sintéticos e reais, e de verificar a mais valia do uso de restrições no agrupamento de dados.

Capítulo 5

Combinação de Agrupamentos de Dados com Restrições

5.1 Introdução

Neste capítulo apresenta-se o conceito de combinação de soluções, isto é, a combinação de vários classificadores na aprendizagem supervisionada e a combinação de agrupamentos de dados na aprendizagem não supervisionada. A combinação de soluções é bastante estudada, tanto na aprendizagem supervisionada como na aprendizagem não supervisionada, com o intuito de aumentar o desempenho dos algoritmos de aprendizagem quando usados individualmente. Com o objectivo de mostrar a mais valia da combinação de soluções, abordam-se sucintamente os tópicos essenciais da Combinação de Classificadores na secção 5.2 e os tópicos principais da Combinação de Agrupamentos de Dados na secção 5.3. Finalmente, na secção 5.4 são propostas quatro abordagens para a combinação de agrupamentos de dados usando restrições de ligações obrigatória e proibida.

5.2 Combinação de Classificadores

A combinação de classificadores tem sido bastante estudada nos últimos anos. O seu grande objectivo é aumentar o desempenho da tarefa de classificação, usando para isso diversos classificadores obtidos por um ou vários algoritmos de classificação.

5.2.1 Problemas da Aplicação de Algoritmos de Classificação Individualmente

A combinação de classificadores tem como objectivo aumentar a exactidão da classificação de dados, tentando resolver os problemas da utilização de classificadores individualmente. Dietterich [25] expôs três tipos de problemas associados à utilização de algoritmos de aprendizagem:

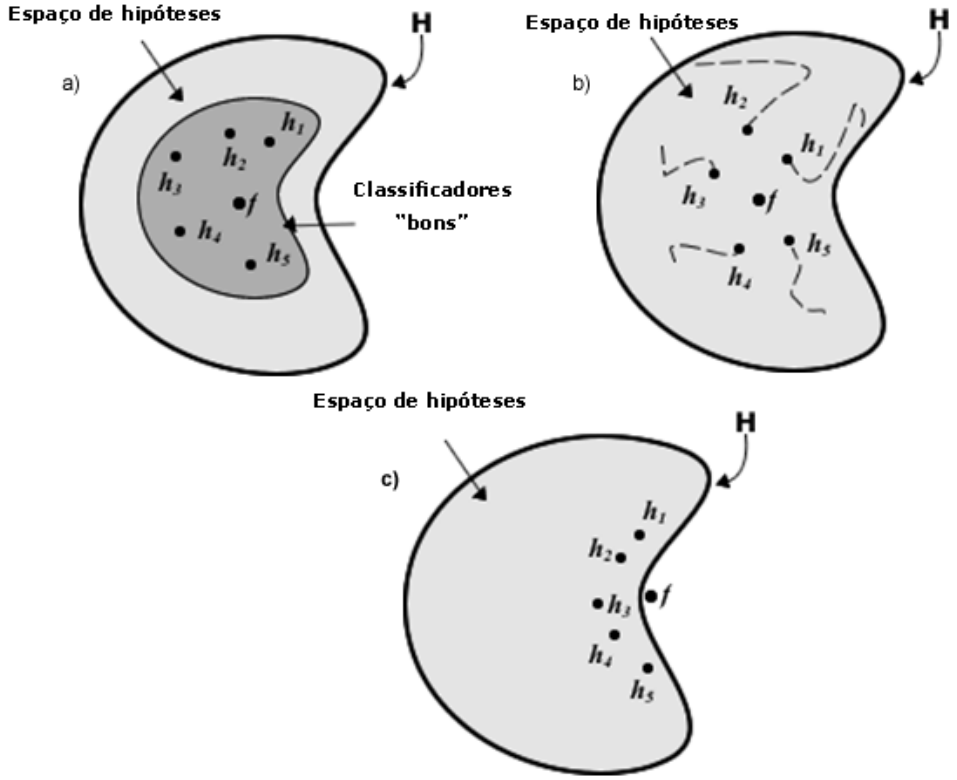

Figura 5.1: Problemas resultantes da aplicação dos algoritmos de classificação individualmente - (a) problema estatístico; (b) problema computacional; (c) problema representativo (adaptado de [25]).

problemas estatísticos, problemas computacionais e problemas representativos. O problema estatístico surge quando o algoritmo de aprendizagem pesquisa um espaço de hipóteses demasiado elevado, tendo em conta o tamanho do conjunto de treino disponível. Sem dados suficientes disponíveis, podem existir hipóteses que sendo substancialmente diferentes atingem a mesma exactidão para o conjunto de treino, tendo o algoritmo de aprendizagem de escolher uma dessas hipóteses. Existe ainda o risco da hipótese escolhida não conseguir classificar correctamente novos exemplos de dados. Uma forma de reduzir este risco consiste numa simples votação de todas as hipóteses [55]. A figura 5.1 a) ilustra o problema estatístico da utilização de apenas um algoritmo de aprendizagem em que o melhor classificador é representado por f e a área cinzenta representa o espaço de hipóteses h_i com bom desempenho. A combinação de vários mínimos locais reduz o risco de se escolher um "mau" mínimo local como hipótese final.

Quando o algoritmo de aprendizagem não garante que encontra a melhor hipótese dentro do espaço de hipóteses estamos na presença de um problema computacional. Nos casos de algoritmos de árvores de decisão ou redes neuronais, a tarefa de encontrar a hipótese que melhor se adequa ao conjunto de dados é computacionalmente intratável, pelo que se recorre ao uso de heurísticas. É comum que as heurísticas encontrem mínimos locais e não o pretendido mínimo global, falhando assim a determinação da melhor hipótese. A figura 5.1 b) ilustra esta situação. O melhor classificador é representado por f e as linhas a tracejado mostram as trajectórias de hipóteses no decorrer do treino do classificador. Analogamente ao problema estatístico, a

combinação de vários mínimos locais pode reduzir o risco de se escolher um mínimo local errado como hipótese final.

Por último, o problema representativo ocorre quando o espaço de hipóteses não inclui hipóteses que sejam boas aproximações da função óptima, tal como ilustrado na figura 5.1 c). A soma ponderada de todas as hipóteses permite por vezes expandir o espaço de funções que podem ser representadas. Assim, através da votação ponderada das hipóteses poderá ser possível que o algoritmo de aprendizagem consiga aproximar com maior exactidão a função óptima.

5.2.2 Abordagens para a Combinação de Classificadores

A combinação de classificadores é uma área que se tem tornado bastante activa nos últimos anos, apesar de ser uma área de investigação com pouco mais de uma década, não existindo ainda consenso sobre a sua taxonomia. Na literatura encontram-se descritas várias abordagens para a combinação de classificadores. Uma das possibilidades é serem agrupadas consoante o seu método de construção [55].

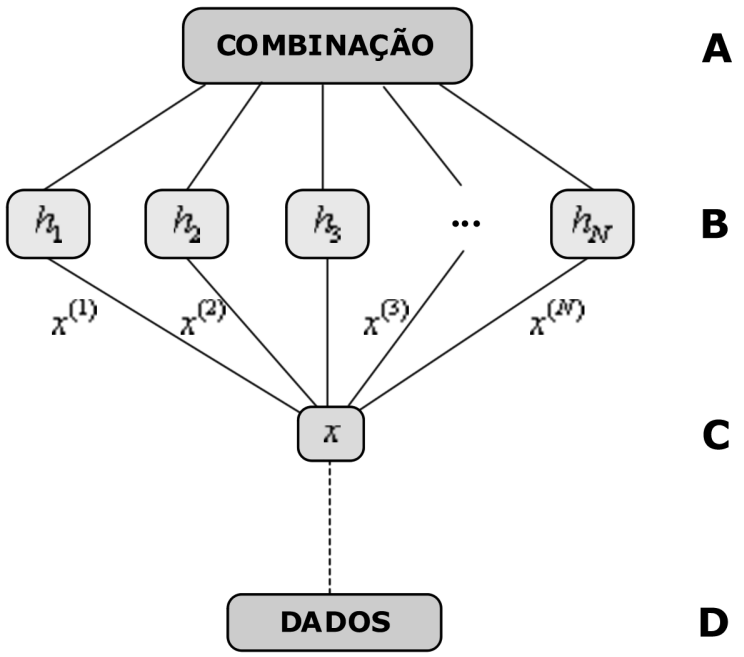

Figura 5.2: Abordagens para a combinação de classificadores - A - nível de combinação; B - nível do classificador; C - nível dos atributos; D - nível de dados.

No diagrama da figura 5.2 apresentam-se quatro diferentes conjuntos de abordagens para construir métodos de combinação de classificadores.

A - Abordagens no Nível de Combinação. Variam na forma de combinação das hipóteses resultantes dos classificadores. As abordagens mais directas usam geralmente um dos seguintes esquemas de votação.

Votação Simples. A classe atribuída por cada classificador é considerada como um voto, tendo cada classificador a mesma importância na classificação. A classe atribuída a

um objecto é simplesmente a classe mais votada entre todos os classificadores [59]. Este tipo de votação é também referida por *votação por maioria* ou SAM (*Select All Majority*).

Votação Ponderada. Contrariamente à votação simples, na votação ponderada cada voto tem um grau de importância diferenciado, geralmente, proporcional ao desempenho de cada classificador (estimativa da generalização) [94]. A motivação é simples: os classificadores com melhor qualidade de generalização devem ter mais importância que os classificadores com pior desempenho.

Maioria Ponderada. A principal diferença entre a abordagem de maioria ponderada e a votação ponderada consiste na forma com que a ponderação de cada voto é obtida, já que na maioria ponderada as ponderações de cada voto são dinâmicas [56]. Nesta abordagem, cada classificador tem inicialmente a mesma importância, isto é, cada voto tem ponderação 1. De seguida, treina-se cada um dos classificadores e caso algum classificador preveja erradamente a classe de um objecto, a sua ponderação (importância na combinação de classificadores) é multiplicado por um valor $0 \leq \alpha \leq 1$. Se $\alpha = 0$, quando um classificador errar a classificação de um objecto, é automaticamente descartado. Se $\alpha > 0$, a ponderação dos classificadores que erram diminui gradualmente, especialmente os classificadores que mais falham.

B - Abordagens no Nível dos Classificadores. Utilizam diversos algoritmos de classificação para gerarem as hipóteses a serem combinadas. As abordagens ao nível dos classificadores abrangem duas grandes categorias:

Combinação de Classificadores Homogéneos. Esta categoria é composta pelas abordagens de combinação de classificadores que apostem num único algoritmo de classificação para gerar as várias hipóteses a combinar. Exemplos desta abordagem são: a combinação de classificadores obtidos pela aprendizagem Bayesiana Ingénua [71]; a combinação de hipóteses obtidas pelo algoritmo de classificação e-GASEN, que é baseado numa rede neuronal de duas camadas [92]; e a combinação de classificadores resultantes do algoritmo de classificação K-vizinho-mais- próximo [94].

Combinação de Classificadores Heterogéneos. Nesta categoria de combinação de classificadores estão incluídos todas as abordagens que utilizem vários tipos de classificadores para formar o conjunto de classificadores. Dois exemplos desses métodos são a Generalização em Pilha [91] e a Meta-Aprendizagem [84].

C - Abordagens no Nível dos Atributos. Utilizam diferentes subconjuntos de atributos como entrada nos algoritmos de classificação. A partir de um único conjunto de dados é possível formar diversos subconjuntos diferentes de atributos do conjunto de dados. Assim, podem-se treinar vários classificadores diferentes para classificar os mesmos dados, variando o subconjunto de atributos usado no treino de cada classificador. O problema

da combinação de classificadores usando subconjuntos de atributos diferentes não exige o desenvolvimento de um esquema especial de combinação dos classificadores. A maior parte dos métodos baseados em votação podem ser aplicados directamente na combinação de classificadores obtidos com subconjuntos de atributos diferentes, já que a decisão final é baseada apenas nos votos de cada classificador, não sendo necessário qualquer acesso aos atributos dos dados.

D - Abordagens no Nível dos Dados. Utilizam subconjuntos diferentes de dados para gerar cada hipótese. Esta categoria agrupa os métodos de combinação de classificadores em que os algoritmos de classificação são treinados várias vezes utilizando um subconjunto diferente dos dados de treino. Este tipo de combinação de classificadores tem um bom desempenho, especialmente quando os algoritmos de classificação utilizados são instáveis, isto é, quando os algoritmos de classificação produzem resultados diferentes com pequenas modificações do conjunto de treino. Entre os métodos desta categoria destacam-se o *Bagging* e o *Boosting* que têm sido bastante estudados e comparados com outros métodos, obtendo geralmente desempenhos superiores [9; 69]. O *Bagging* [17] é um método de combinação de classificadores que combina vários classificadores construídos a partir réplicas do conjunto de treino. As réplicas são obtidas através de amostragem com reposição do conjunto de treino. Depois de se treinarem todos os classificadores, os objectos são atribuídos à classe mais votada. O *Boosting* [37] sugere que a combinação de classificadores *fracos*, isto é, classificadores com desempenhos pouco superiores à escolha aleatória, pode originar um classificador *forte*, com um desempenho de grande qualidade. Nesta abordagem, cada classificador é construído sequencialmente através de amostragem de dados com reposição, considerando o desempenho do classificador obtido na iteração anterior. A distribuição da probabilidade de um objecto ser seleccionado numa amostra de dados não é uniforme, contrariamente ao que acontece no *Bagging*. Cada objecto do conjunto de treino tem uma ponderação associada, indicando a probabilidade desse objecto ser incluído na amostra do conjunto de treino usada para treinar o próximo classificador. Os objectos mais difíceis de classificar, aqueles em que os classificadores erram mais vezes na sua classificação, têm uma ponderação mais elevada enquanto que os objectos mais fáceis de classificar têm uma ponderação inferior. Assim, nos treinos sucessivos de classificadores é dada maior importância a objectos cuja classificação é difícil.

5.3 Combinação de Agrupamentos de Dados

A combinação de agrupamentos de dados surgiu na última década com o intuito de melhorar a robustez e qualidade do agrupamento de dados, reutilizar soluções e agrupar dados de forma distribuída. Nesta secção são apresentadas as grandes vantagens da combinação de agrupamentos de dados e indicadas várias abordagens para a combinação de agrupamentos de dados.

5.3.1 Vantagens da Combinação de Agrupamentos de Dados

Com a combinação de agrupamentos de dados pretende-se produzir um agrupamento de um conjunto de dados, usando para isso agrupamentos de dados obtidos por algoritmos de agrupamentos (aplicados individualmente), sem que para isso seja (geralmente) necessário aceder aos atributos do conjunto de dados. De seguida, apresentam-se a principais vantagens da combinação de agrupamentos de dados [27].

Aumento da qualidade do agrupamento de dados. A combinação de vários agrupamentos de dados, num agrupamento de dados final, origina normalmente um agrupamento do conjunto de dados com melhor qualidade e robustez que os agrupamentos obtidos com a aplicação de um algoritmo de agrupamento de dados individualmente.

Agrupamento de dados fisicamente distribuídos. Actualmente, as organizações adquirem cada vez mais informação, armazenando-a em vários pontos devido a motivos operacionais, organizacionais e à tendência crescente de se realizar localmente o tratamento da informação. Os algoritmos de agrupamento de dados tradicionais assumem que o conjunto de dados a ser tratado encontra-se centralizado, o que muitas vezes é uma assumpção inválida, pelo que é necessário transferir toda a informação distribuída para o mesmo local. No entanto, a centralização dos dados pode ser muito difícil de realizar, ou mesmo impossível, devido aos custos de processamento, de armazenamento e da transferência dos dados, existindo também outro tipo de restrições, tais como, a propriedade dos dados, a legislação e a segurança.

Reutilização de agrupamentos de dados. Outra vantagem da combinação de agrupamentos de dados consiste na reutilização de agrupamentos de dados já existentes. Para um determinado conjunto de dados, podem já existir vários agrupamentos pelo que pode ser vantajoso combinar esse agrupamentos num único agrupamento de dados ou usá-los conjuntamente com novos agrupamentos de dados com o intuito de influenciar o agrupamento de dados final.

Aceleração do processo de agrupamento de dados. Por vezes, dependendo dos algoritmos de agrupamento escolhidos e da abordagem de combinação, é possível acelerar o agrupamento de dados desde que seja usado paralelismo no agrupamento de amostras do conjunto de dados original.

Agrupamento de dados heterogéneos. Existem situações em que o conjunto de dados é caracterizado por vários subconjuntos de atributos relacionados entre si, não existindo qualquer correlação entre atributos de subconjuntos diferentes. A aplicação de um algoritmo de agrupamento para produzir um único agrupamento de dados não é normalmente eficaz nestas situações. No entanto, produzir vários agrupamentos de dados (por exemplo,

um agrupamento para cada subconjunto de atributos) permite que sejam encontrados grupos em subespaços dos atributos de dados, podendo a combinação desses agrupamentos levar a um agrupamento de dados final com boa qualidade.

Tendo sido apontadas as principais vantagens da combinação de agrupamentos de dados, na próxima secção apresentam-se várias abordagens para a realizar.

5.3.2 Abordagens para a Combinação de Agrupamentos de Dados

Existem várias formas para realizar a combinação de agrupamentos de dados, tal como apresentado na figura 5.3. Como se pode verificar, as abordagens podem ser categorizadas considerando a forma como os conjuntos de agrupamentos a combinar são produzidos e o esquema de combinação desses agrupamentos, isto é, a função de consenso [27]. De seguida, são explicados os vários tipos de métodos de construção de conjunto de agrupamento de dados e de funções de consenso ilustrados na figura 5.3.

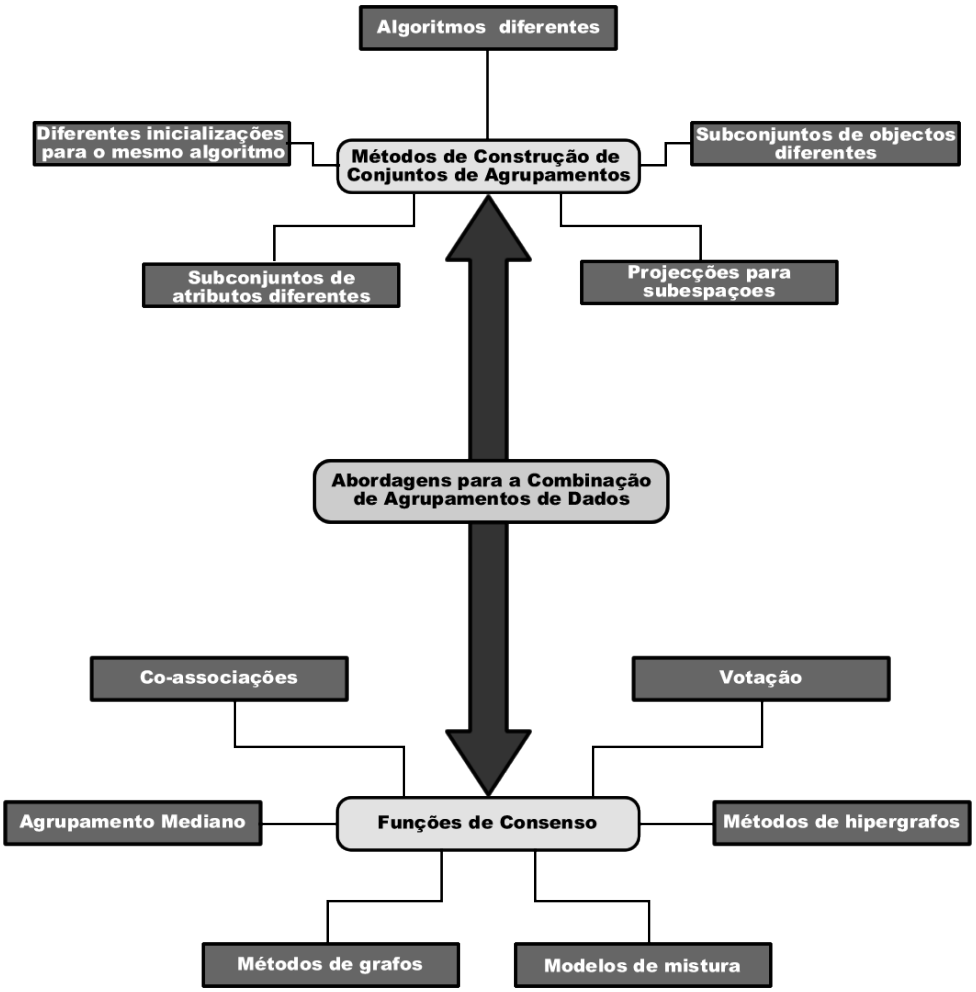

Figura 5.3: Abordagens para a combinação de agrupamentos - Métodos de construção de conjuntos de agrupamentos e funções de consenso.

5.3.2.1 Métodos de Construção de Conjuntos de Agrupamentos de Dados

O método de construção do conjunto de agrupamentos de dados define o modo como são gerados os agrupamentos de dados a combinar. Neste passo, a criação de diversidade no conjunto de agrupamentos origina geralmente agrupamentos de dados finais de qualidade superior, pois exite uma relação entre qualidade do agrupamento de dados final e a diversidade dos agrupamentos de dados que lhe deram origem [41]. De seguida apresentam-se vários métodos para a construção de conjuntos de agrupamentos, podendo estes ser utilizados separada ou conjuntamente com outros métodos de construção.

Diferentes algoritmos de agrupamento. O conjunto de agrupamentos de dados pode ser produzido usando apenas um algoritmo de agrupamento. No entanto, a aplicação de diferentes algoritmos de agrupamento de dados cria uma maior diversidade no conjunto de agrupamentos o que pode influenciar positivamente a qualidade do agrupamento de dados final [28].

Inicializações diferentes para o mesmo algoritmo. Mesmo que se utilize apenas um algoritmo de agrupamento de dados, na construção do conjunto de agrupamentos, é possível fazer com que este possua uma grande diversidade. Uma das formas de o realizar consiste na aplicação de algoritmos de agrupamento com diferentes parâmetros (por exemplo, variar o número de grupos) e diversas inicializações (no caso do K-médias, definir centróides iniciais distintos em cada execução) [36]. Outra forma, resume-se ao uso de versões relaxadas do mesmo algoritmo de agrupamento, com o intuito de se produzir agrupamentos de dados correspondentes a óptimos locais das versões relaxadas dos algoritmos [79].

Subconjuntos de objectos diferentes. Na construção do conjunto de agrupamentos não é necessário que todos os objectos do conjunto de dados tenham sido incluídos na produção de todos os agrupamentos. Na realidade, pode ser bastante proveitoso usar subconjuntos de objectos diferentes para produzir cada agrupamento de dados. Minaei-Bidgoli *et al.* [61] propuseram a utilização de amostragem com reposição para a construção de conjuntos de agrupamentos com base na seguinte noção: a utilização de subconjuntos de dados diferentes na produção de cada agrupamento deverá originar grupos diferentes. Com a combinação destes grupos, espera-se que o agrupamento de dados final seja mais estável que os agrupamentos produzidos pelos algoritmos de agrupamento individualmente. Assim, na geração de cada agrupamento, em vez de se usar a totalidade do conjunto de dados, uma amostra com o mesmo número de objectos do conjunto de dados é construída, escolhendo-se aleatoriamente e com reposição objectos do conjunto de dados original. Note-se que, devido à amostragem ser realizada com repetição, a amostra de dados contém geralmente muitos objectos repetidos.

Outra forma de utilizar subconjuntos de objectos diferentes em cada agrupamento consiste no uso de amostragem sem reposição [28]. Neste caso, as amostras de dados usadas para

produzir cada agrupamento de dados, contêm apenas uma fracção dos objectos do conjunto de dados originais, sendo os objectos de cada amostra de dados todos diferentes.

O método de combinação de agrupamentos proposto por Topchy *et al.* [80] emprega amostragem com reposição para gerar cada agrupamento de dados, mas usando uma abordagem adaptativa, inspirada no sucesso dos algoritmos de *Boosting* [37] da classificação de dados. Neste método, o conjunto de agrupamentos é construído com amostras de dados obtidas sequencialmente. Contrariamente ao método de Minaei-Bidgoli *et al.*, cada objecto de dados tem associada a probabilidade de ser incluído em cada amostra de dados sendo esta actualizada dinamicamente, de forma a que, os objectos que sejam mais difíceis de agrupar tenham uma maior probabilidade de ser escolhidos.

Subconjuntos de atributos diferentes. Este tipo de construção de conjuntos de agrupamentos de dados usa apenas um subconjunto dos atributos de dados para gerar cada agrupamento de dados. Cada subconjunto de atributos pode ser considerado uma vista parcial do conjunto de dados, pelo que a combinação dos vários agrupamentos de dados é encarada como a agregação de vistas diferentes sobre o conjunto de dados. O uso de subconjuntos de atributos diferentes permite também o agrupamento de dados quando estes se encontram fisicamente distribuídos, reduz a quantidade de memória necessária para realizar o agrupamento de dados, permite o tratamento de conjuntos de dados heterogéneos, podendo ainda aumentar a qualidade do agrupamento de dados final devido à agregação de várias vistas diferentes do conjunto de dados.

Projecções para subespaços. O último tipo de construção de conjuntos de agrupamentos consiste na projecção do espaço original dos atributos de dados num espaço de dimensionalidade reduzida. Esta redução do número de atributos tem dois objectivos: evitar o efeito de possíveis atributos ruidosos, ou mesmo irrelevantes, que podem induzir o algoritmo de agrupamento em erro; e proporcionar um melhor agrupamento de dados, visto que os objectos tendem a estar dispersos num conjunto de dados de elevada dimensionalidade. Alguns exemplos de métodos usados na combinação de agrupamentos para projectar o espaço de atributos original noutro de dimensionalidade reduzida são: a análise dos componentes principais [70]; a projecção aleatória [31], em que os atributos originais de dados são linearmente combinados, sendo as ponderações de cada atributo original obtidas aleatoriamente; e a abordagem de Topchy *et al.* [79] que se baseia na construção de conjuntos de agrupamentos *fracos*, obtidos pela projecção do espaço de atributos para apenas uma dimensão ou pela definição de hiperplanos aleatórios no espaço de atributos que determinam se um objecto x_i se encontra, ou não, no mesmo grupo que outro objecto x_j.

O passo seguinte à construção do conjunto de agrupamentos de dados consiste na combinação desses mesmos agrupamentos em apenas um agrupamento de dados final, que se espera ser de qualidade superior aos agrupamentos que o originaram. De seguida, são apresentadas

várias abordagens para a produção de um agrupamento de *consenso* a partir de um conjunto de agrupamentos de dados.

5.3.2.2 Funções de Consenso

Uma função de consenso f consiste numa função que produz um agrupamento de dados final P^* (agrupamento de consenso), combinando um conjunto de N agrupamentos de dados $\mathcal{P} = \{P^1, \cdots, P^N\}$, obtidos através da aplicação de um ou vários algoritmos de agrupamento Alg_i a um conjunto de n objectos de dados, $\mathcal{X} = \{x_1, \cdots, x_n\}$, tal como exemplificado na figura 5.4.

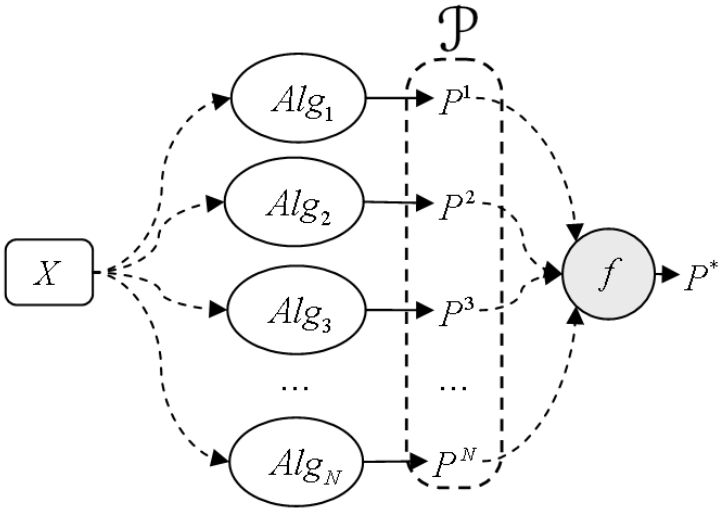

Figura 5.4: Função de Consenso - Combinação de N agrupamentos de dados P^i num agrupamento de dados final P^* (retirado de [27]).

Formalmente, uma função de consenso $f : \{P^1, \cdots, P^N\} \rightarrow P^*$ mapeia o conjunto de agrupamentos $\mathcal{P}$ num agrupamento de consenso P^*, tal que:

$$f : \left\{\{C_1^1, \cdots, C_{K^1}^1\}, \cdots, \{C_1^i, \cdots, C_{K^i}^i\}, \cdots, \{C_1^N, \cdots, C_{K^N}^N\}\right\} \rightarrow \{C_1^*, \cdots, C_{K^*}^*\} \qquad (5.1)$$

De seguida, são descritas várias abordagens de funções de consenso, sendo apontadas as abordagens mais representativas de cada tipo.

Abordagens de votação. As abordagens de votação são as abordagens mais usadas na combinação de classificadores. Tal como mencionado na subsecção 5.2.2, cada classificador vota numa classe para um objecto x_i, sendo x_i classificado com a classe mais votada. As abordagens de votação na combinação de agrupamentos de dados baseiam-se na mesma ideia. Contudo, na aprendizagem não supervisionada o problema é bem mais complexo: o grupos a que pertence cada objecto de dados é desconhecido. Por este motivo, é necessário realizar a correspondência entre cada um dos grupos C_l^j existentes no agrupamento de dados $P^j = \{C_1^j, \cdots, C_{K^j}^j\}$ com cada grupo C_m^i dos restantes agrupamentos de dados

$P^i = \{C_1^i, \cdots, C_{K^i}^i\}$, o que se torna computacionalmente intratável quando o número de agrupamentos de dados a combinar e o número de grupos em cada agrupamento não é trivial. Por este motivo, é necessária a utilização de heurísticas que aproximem a melhor correspondência entre grupos do conjunto de agrupamentos. Note-se que a correspondência entre grupos de apenas dois agrupamentos de dados é um problema bastante mais simples, sendo eficientemente calculada usando o método Húngaro [54].

Um exemplo de uma função de consenso baseada em votação é o algoritmo *Voting-Merging* [26]. Este algoritmo consiste nos três seguintes procedimento:

Procedimento de voto. Neste procedimento, são iterativamente combinados o agrupamento de dados corrente (inicialmente P^1, o primeiro agrupamento de $\mathcal{P}$) com outro agrupamento de dados $P^i \in \mathcal{P}$. No *Voting-Merging*, a correspondência entre grupos é realizada encontrando o par de grupos (um grupo pertencente a cada agrupamento de dados) cuja cardinalidade da intersecção de objectos é a maior. Em seguida, esses grupos são descartados e o processo é repetido até que exista correspondência entre todos os grupos. Após ter sido realizada a correspondência entre grupos, é realizada a votação. Estes passos são repetidos até que todos os agrupamentos de dados tenham sido votados.

Agrupamento resultante do procedimento de voto. Quando todos os agrupamentos de dados tiverem sido processados, a fracção de vezes em que cada objecto de dados foi associado a cada grupo é conhecida pelo que é possível atribuir um grupo a cada objecto de dados. No entanto, o conjunto de agrupamentos usado pelo *Voting-Merging* é construído de forma a que o número de grupos dos agrupamentos a combinar seja superior ao número de grupos do agrupamento de dados final, pelo que se segue ainda um procedimento de convergência dos grupos.

Procedimento de convergência. Finalmente, no procedimento de convergência são fundidos pares ou *correntes* de grupos até que se encontre um critério de paragem. Os grupos são fundidos tendo em conta a cardinalidade da intersecção da fracção de objectos existentes em cada grupo.

Outro exemplo de um método baseado em votação é o BagClust1 [29]. Este método constrói o conjunto de agrupamentos usando amostragem com reposição e utiliza um agrupamento de referência (produzido usando todos os objectos do conjunto de dados) que serve apenas para realizar a correspondência entre todos os grupos dos agrupamentos de dados a combinar. Pelo facto de nem todos os objectos de dados serem incluídos em todos os agrupamentos de dados a combinar, o total de votos de cada objecto tem de ser normalizado tendo em conta o número de vezes que este é seleccionado para uma amostra de dados. Cada objecto é atribuído ao grupo em que foi mais vezes votado.

Abordagens baseadas em co-associações. Os métodos de combinação de agrupamentos baseados em co-associações constroem uma matriz simétrica e positiva $co\_assoc \in \mathbb{R}_0^{+(n\times n)}$ que contabiliza a frequência com que cada par de objectos (x_i, x_j) foi agrupado conjuntamente no conjunto de agrupamentos $\mathcal{P}$. Como apenas são consideradas relações entre pares de objectos, não existe o problema da correspondência entre grupos de agrupamentos diferentes, existente nas abordagens baseadas em votação. O agrupamento de dados de consenso é obtido usando apenas a matriz *co_assoc*. De seguida apresentam-se exemplos de métodos de combinação de agrupamentos que usam o conceito das co-associações entre objectos.

Fred propôs o método Acumulação de Evidências [35] (*Evidence Accumulation Clustering* - EAC) baseando-se na ideia de que se um par de objectos pertence ao mesmo grupo *natural*, o par de objectos será, na maior parte das vezes, agrupado no mesmo grupo no conjunto de agrupamentos de dados a combinar. A fracção de vezes que dois objectos, x_i e x_j, são associados ao mesmo grupo, no conjunto dos N agrupamentos de dados a combinar, é guardada na entrada co_assoc_{ij} da matriz de co-associações. Esta matriz representa uma nova medida de similaridade entre objectos de dados, pelo que pode ser usada por um algoritmo de agrupamento de dados para se produzir um agrupamento de dados de consenso. O algoritmo 5.1 resume o método de Acumulação de Evidências.

Algoritmo 5.1: Acumulação de Evidências

Entrada: $\mathcal{P} = \{P^1, \cdots, P^N\}$ - Conjunto de N agrupamentos de dados, $Rest_=$ - Restrições de ligação obrigatória, $Rest_\neq$ - Restrições de ligação proibida
Saída: P^* - Agrupamento de dados de consenso

```
Criar uma matriz de co-associações nula de tamanho n × n, co_assoc, em que n é o número de objecto ;
/*Calcular co-associações entre pares de objectos*/;
para cada P^l ∈ P faça
    /*Actualizar co-associações entre pares de objectos colocados no mesmo grupo no agrupamento P^l*/;
    para cada C_k^l ∈ P^l faça
        para cada (x_i, x_j) ∈ C_k^l faça
            co_assoc_ij = co_assoc_ij + 1/N;
        fim
    fim
fim
Aplicar um algoritmo de agrupamento de dados, usando como entrada co_assoc para se obter o agrupamento de consenso P*;
```

O método de combinação de agrupamentos Clusterfusion [50] difere substancialmente do método de Acumulação de Evidências na forma como o agrupamento de dados de consenso

é obtido. Com base na matriz de co-associações (não sendo esta normalizada pelo número de agrupamentos a combinar, N), é criada uma lista L contendo todos os pares de objectos agrupados no mesmo grupo em todos os N agrupamentos de dados. De seguida, é criado um conjunto vazio de grupos P e o primeiro par de objectos existente em L é adicionado a P, formando o primeiro grupo. A lista L é então percorrida da seguinte forma: se um objecto do par corrente pertencer a um dos grupos C_k de P, o outro objecto do par é adicionado a C_k; caso contrário, é criado um novo grupo grupo C_{k+1} com os dois objectos. O agrupamento de dados de consenso é obtido quando se verificarem todos os pares de objectos contidos em L.

O BagClust2 [29] é também um método de agrupamento baseado em co-associações, sendo o conjunto de agrupamentos de dados produzido usando amostragem com reposição, da mesma forma que o BagClust1. Por este motivo, a normalização da matriz de co-associações não pode ser realizada do mesmo modo que no método de Acumulação de Evidências (divisão por N), pois nem todos os objectos de dados aparecem em cada agrupamento de dados a combinar. Neste método, cada entrada da matriz de co-associações, co_assoc_{ij}, é obtida calculando o número de vezes que os objectos x_i e x_j foram associados ao mesmo grupo dividido pelo número de vezes em que os dois objectos foram incluídos simultaneamente nos agrupamentos de dados. Após co_assoc ter sido calculada, é criada uma matriz de dissimilaridades D de tamanho $n \times n$ em que cada entrada é calculada por $D_{ij} = 1 - co\_assoc_{ij}$. Finalmente, o agrupamento de consenso é obtido aplicando uma algoritmo de agrupamento de dados baseado em dissimilaridades entre pares de objectos.

Agrupamento Mediano. As abordagens de agrupamento de dados baseadas no *agrupamento mediano* procuram um agrupamento de consenso P^* que maximize a semelhança entre P^* e todos os agrupamentos de dados $P^i \in \mathcal{P}$. A semelhança entre dois agrupamentos de dados P_i e P_j pode ser calculada de diferentes formas.

A Informação Mútua quadrática (IM^2) [77] pode ser usada para medir a similaridade entre agrupamentos de dados. No entanto, procurar exaustivamente o agrupamento de dados P^* que maximiza a IM^2 média entre todos os pares de agrupamentos $(P^*, P^i), \forall P^i \in \mathcal{P}$ é computacionalmente intratável. Mirkin provou [62] que a minimização da Informação Mútua quadrática entre cada agrupamento de dados a combinar e P^* é equivalente à minimização do erro quadrático, se o conjunto de agrupamentos $\mathcal{P}$ for transformado num novo conjunto de dados binário $\mathcal{X}'$ com $\sum_{i=1}^{N} K^i$ atributos, correspondendo cada atributo a um grupo contido no conjunto de agrupamentos $\mathcal{P}$, do seguinte modo: se os objectos tiverem sido associados ao grupo C_k, então o valor para o atributo correspondente a C_k toma o valor 1; caso contrário, o valor para o atributo toma o valor 0; finalmente, o novo conjunto de dados $\mathcal{X}'$ é estandardizado de forma a que cada atributo tenha valor médio 0. Com base nesta constatação, Topchy *et al.* [79] propuseram que o algoritmo de agrupamento K-médias fosse usado como função de consenso, já que este minimiza o erro

quadrático existente num conjunto de dados, transformando o conjunto de agrupamentos $\mathcal{P}$ num novo conjunto de dados $\mathcal{X}'$ tal como anteriormente explicado.

Jouve e Nicoloyannis [44] propuseram um método de combinação de agrupamentos de dados baseado no algoritmo de agrupamento de dados categóricos KEROUAC [45]. O KEROUAC minimiza uma função de erro denominada *New Condorcet Criterion* que contabiliza o número de similaridades entre objectos pertencentes a grupos diferentes e o número de dissimilaridades entre objectos que pertencem ao mesmo grupo. O método de combinação proposto constrói uma nova representação do conjunto de agrupamentos de dados, transformando cada agrupamento P^i num atributo da nova representação do conjunto de dados, $\mathcal{X}'$. O algoritmo de agrupamento de dados KEROUAC é aplicado ao novo conjunto de dados $\mathcal{X}'$, sendo o agrupamento resultante o agrupamento de dados final, P^*.

Métodos baseados em hipergrafos. Nesta abordagem, o conjunto de agrupamentos de dados é mapeado num hipergrafo, transformando a combinação de agrupamentos num simples problema de partição de grafos. Strehl e Gosh [77] propuseram três métodos de combinação de agrupamentos com base num hipergrafo construído da seguinte forma: para cada agrupamento de dados $P^i \in \mathcal{P}$ é construída uma matriz binária H^i, com uma coluna para cada grupo $C_k^i \in P^i$ (representando uma hiperaresta) e uma linha para cada um dos n objectos de dados (representando os vértices); para cada vértice, cada hiperaresta h_a tem o valor 0 ou 1 associado, indicando se cada objecto foi (valor 1) ou não (valor 0) associado ao grupo representado por h_a; concatenando todas as matrizes H_i, correspondentes a cada agrupamento de dados a combinar, obtém-se a matriz de adjacência do hipergrafo, $\mathcal{H} = \{H_1, \cdots, H_N\}$. A figura 5.5 mostra um exemplo da construção de um hipergrafo, com base num conjunto de agrupamentos de dados.

O método CSPA (*Cluster-based Similarity Partition Algorithm*) é o primeiro dos métodos propostos por Strehl e Gosh. A ideia deste método consiste em transformar o hipergrafo $\mathcal{H}$ numa matriz de similaridades S entre pares de objectos, à semelhança das abordagens baseadas em co-associações, e em seguida realizar a partição do grafo representado pela matriz S em K componentes, usando um algoritmo de partição de grafos, como é exemplo o METIS [48]. A matriz de similaridades S indica a frequência com que cada par de objectos foram associados ao mesmo grupo, no total dos N agrupamentos de dados a combinar, e é calculada aplicando a seguinte equação:

$$S = \frac{1}{N} H H^{\mathrm{T}} \tag{5.2}$$

em que H^{T} representa a transposta de H.

Outro dos métodos propostos é o HGPA (*Hyper-Graph Partitioning Algorithm*), que realiza directamente a partição do hipergrafo definido por $\mathcal{H}$ usando o algoritmo de partição

	P^1	P^2	P^3			H_1			H_2			H_3	
						h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8
x_1	1	3	2		v_1	1	0	0	0	0	1	0	1
x_2	1	3	2		v_2	1	0	0	0	0	1	0	1
x_3	2	3	?	$\Leftrightarrow$	v_3	0	1	0	0	0	1	0	0
x_4	2	1	1		v_4	0	1	0	1	0	0	1	0
x_5	2	1	1		v_5	0	1	0	1	0	0	1	0
x_6	3	2	?		v_6	0	0	1	0	1	0	0	0
x_7	3	2	1		v_7	0	0	1	0	1	0	1	0
x_8	3	2	?		v_8	0	0	1	0	1	0	0	0

Figura 5.5: Construção de um hipergrafo com base num conjunto de agrupamentos de dados - Cada grupo existente no conjunto de agrupamentos é representado por um vector h_i e cada agrupamento por um conjunto de vectores H_j.

de hipergrafos HMETIS [46] para obter o agrupamento de dados de consenso. O método HGPA baseia-se na ideia de que as ligações entre os grupos do conjunto de agrupamentos são indicadores de ligações fortes entre os objectos que lhes estão associados. O agrupamento de consenso é determinado eliminando sucessivamente hiperarestas do hipergrafo $\mathcal{H}$ até que restem apenas K componentes de vértices isolados, formando-se grupos com os objectos correspondentes aos vértices de cada componente isolado.

O último método de combinação de agrupamentos proposto por Strehl e Gosh é o MCLA (*Meta-CLustering Algorithm*). A ideia deste método consiste no agrupamento e desagregação de hiperarestas de $\mathcal{H}$ que se encontrem relacionadas, atribuíndo-se cada objecto à hiperaresta a que foi mais vezes atribuído.

Métodos baseados em grafos. Os métodos baseados em grafos mapeiam o conjunto de agrupamentos a combinar $\mathcal{P}$ num grafo, sendo o agrupamento de consenso, geralmente, obtido com a partição desse grafo em K componentes. Fern e Brodley [32] propuseram três métodos de combinação de agrupamentos baseados em grafos, em que cada método corresponde a uma formulação diferente da construção do grafo. O primeiro método, denominado IBGF (*Instance-Based Graph Formulation*), constrói um grafo em que são modeladas as relações entre cada par de objectos do conjunto de dados, seguindo as ideias da abordagem baseada em co-associações. O grafo é composto por n vértices, correspondendo cada um a um objecto de dados. Cada vértice encontra-se ligado a todos os outros vértices por arcos ponderados, em que a ponderação de cada arco corresponde à fracção de vezes em que os dois objectos foram associados ao mesmo grupo no conjunto de agrupamentos $\mathcal{P}$.

O agrupamento de consenso é obtido efectuando a partição do grafo em K componentes. Este método é equivalente ao método CSPA, apresentado anteriormente.

O segundo método proposto é o CBGF (*Cluster-Based Graph Formulation*) e, contrariamente ao método anteriormente apresentado, representa as relações entre os vários grupos existentes em $\mathcal{P}$. Para isso, é construído um grafo em que os vértices correspondem aos grupos do conjunto de agrupamentos de dados, existindo um arco entre todos os pares de vértices do grafo, sendo cada arco ponderado usando a medida de Jaccard. Formalmente, as ponderações dos arcos são representadas numa matriz de adjacências A, em que a similaridade entre dois vértices, isto é, dois grupos C_i^l e C_j^m é calculada por

$$A_{ij} = \frac{|C_i^l \cap C_j^m|}{|C_i^l \cup C_j^m|} \tag{5.3}$$

Em seguida, é realizada a partição do grafo em K componentes, correspondendo cada componente a apenas um agrupamento de grupos, e não de objectos, sendo cada componente considerada como um metagrupo. Para se determinar o agrupamento de consenso é necessário associar cada objecto ao metagrupo a que foi mais vezes atribuído.

O método HBGF (*Hybrid Bipartite Graph Formulation*), o último método de combinação de agrupamentos de dados proposto por Fern e Brodley, tem como objectivo mapear num grafo, simultaneamente, as similaridades entre objectos e as similaridades entre grupos existentes nos N agrupamentos de dados a combinar $\{P^1, \cdots, P^N\}$. O método HBGF constrói um grafo com vértices para representar todos os objectos e grupos do conjunto de agrupamentos, existindo arestas apenas entre vértices correspondentes a pares objecto/grupo. Se a aresta relacionar um objecto x_i a um grupo C_j^l em que $x_i \in C_j^l$, a aresta tem ponderação 1. Caso contrário, a aresta tem ponderação 0. Após se ter construído o grafo, o agrupamento de dados de consenso é obtido com a partição deste em K componentes isolados, formando cada conjunto de objectos existente em cada componente, um grupo do agrupamento de consenso.

Al-Razgan e Domeniconi [2] propuseram também dois métodos de combinação baseados em grafos, tendo a particularidade de se especializarem no agrupamento de dados produzidos pelo algoritmo de agrupamento de dados LAC [19] (*Locally Adaptive Clustering*). O LAC é um algoritmo de agrupamento de partição que forma grupos em subespaços dos atributos de dados, ponderando diferentemente cada atributo de dados no cálculo da medida de distância. Os métodos WSPA (W*eighted Similarity Partitioning Algorithm*) e WBPA (*Weighted Bipartite Partitioning Algorithm*) seguem as mesmas ideias que os algoritmos CBGF e HBGF, respectivamente, inovando na forma como as ponderações das arestas são calculadas, pois é também utilizada informação relativa às ponderações dos atributos em cada grupo do conjunto de agrupamentos $\mathcal{P}$.

Abordagens baseadas em modelos de mistura. Estas abordagens de combinação de agrupamentos criam um modelo de mistura para representar o conjunto de agrupamentos de dados $\mathcal{P}$. Topchy *et al.* [81] propuseram um modelo probabilístico para encontrar o agrupamento de consenso no espaço dos grupos do conjunto de agrupamentos, transformando o problema da combinação de agrupamentos num problema de máxima verosimilhança. A função de verosimilhança é optimizada nos parâmetros de uma distribuição de mistura finita, multivariada e multinomial usando o algoritmo EM [24] (*Expectation-Maximization*). Com o modelo proposto, pretende-se optimizar a função de verosimilhança, encontrando os parâmetros que definem cada componente de distribuição, que representam cada grupo do agrupamento de consenso.

5.4 Combinação de Agrupamentos de Dados com Restrições

Nesta secção, são propostos quatro métodos para a combinação de agrupamentos de dados com restrições: na subsecção 5.4.1, é apresentado um método que consiste numa simples modificação do método de Acumulação de Evidências e é denominado *Constrained Evidence Accumulation Clustering* (CEAC); na subsecção 5.4.2, o método CEAC é expandido, usando uma abordagem adaptativa para a construção do conjunto de agrupamentos de dados baseada em *Boosting* [37]; na subsecção 5.4.3, é apresentado um método de combinação de agrupamentos que usa o algoritmo COP-COBWEB como função de consenso; e, finalmente, na subsecção 5.4.4, é proposto um algoritmo genético para optimizar uma função-objectivo baseada na Medida de Consistência dos Grupos [27].

5.4.1 CEAC

A primeira proposta para a combinação de agrupamentos de dados consiste numa simples modificação do método de Acumulação de Evidências [35], previamente apresentado no tópico 5.3.2.2. Esta modificação consiste na utilização de um algoritmo de agrupamento hierárquico de dados com restrições para se extrair da matriz de co-associações *co_assoc* o agrupamento de dados de consenso P^*. Este método é denominado *Constrained Evidence Accumulation Clustering* (CEAC).

Tal como o método de Acumulação de Evidências, o método CEAC mapeia o conjunto de N agrupamentos de dados $\mathcal{P} = \{P^1, \cdots, P^N\}$ numa matriz de co-associações *co_assoc* de tamanho $n \times n$, tendo em conta a frequência com que os pares de objectos do conjunto de dados $(x_i, x_j) \in \mathcal{X} = \{x_1, \cdots, x_n\}$ são agrupados no mesmo grupo, no total dos N agrupamentos de dados que se pretendem combinar. A matriz de co-associações *co_assoc* define uma nova medida de similaridade entre os pares de objectos, sendo a co-associação entre dois objectos de dados, x_a e x_b, calculada por:

$$co\_assoc_{ab} = \frac{\sum_{i=1}^{N} I(l_a^i = l_b^i)}{N} \tag{5.4}$$

em que $I(\cdot)$ devolve 1 caso a expressão seja verdadeira e 0 no caso contrário. l_a^i e l_b^i são os rótulos dos objectos x_a e x_b no i-ésimo agrupamento de dados, P^i.

No método de Acumulação de Evidências, após o cálculo da matriz de co-associações co_assoc, é aplicado um algoritmo de agrupamento de dados para se obter o agrupamento de dados final P^*. No entanto, quando existem relações entre objectos do conjunto de dados, o algoritmo de agrupamento deverá incorporar restrições de ligações obrigatória e proibida para que as restrições sejam totalmente satisfeitas. Na avaliação do método *CEAC*, realizada na secção 6.6, são utilizados o algoritmo de Ligação Completa Restringida (subsecção 4.5.1) e uma versão modificada do algoritmo Ligação Simples, em que, os objectos restringidos com ligações obrigatórias são inicialmente agrupados no mesmo grupo e só é possível juntar pares de grupos cujos objectos não tenham ligações proibidas entre si. O pseudo-código do método CEAC é apresentado no algoritmo 5.2.

Algoritmo 5.2: CEAC

Entrada: $\mathcal{P} = \{P^1, \cdots, P^N\}$ - Conjunto de N agrupamentos de dados, $Rest_=$ - Restrições de ligação obrigatória, $Rest_{\neq}$ - Restrições de ligação proibida

Saída: P^* - Agrupamento de dados de consenso

```
Criar uma matriz de co-associações nula de tamanho n × n, co_assoc, em que n é o número de objecto ;
/*Calcular co-associações entre pares de objectos*/;
para cada P^l ∈ P faça
    /*Actualizar co-associações entre pares de objectos colocados no mesmo grupo no agrupamento P^l*/;
    para cada C_k^l ∈ P^l faça
        para cada (x_i, x_j) ∈ C_k^l faça
            co_assoc_ij = co_assoc_ij + 1/N;
        fim
    fim
fim
Aplicar um algoritmo de agrupamento hierárquico de dados com restrições à matriz de co-associações co_assoc para se obter o agrupamento de consenso P*;
```

Note-se que na construção do conjunto de agrupamentos de dados, os agrupamentos de dados podem ser produzidos usando qualquer algoritmo de agrupamento de dados, com ou sem o uso de restrições, pois as relações entre objectos são impostas através do algoritmo de agrupamento hierárquico de dados com restrições.

5.4.2 CEAC*Boost*

O método CEAC*Boost* é uma expansão do método CEAC que usa uma abordagem adaptativa, inspirada nos algoritmos de *Boosting* [35], na construção do conjunto de agrupamentos de dados a combinar. Contrariamente ao CEAC, em que o conjunto de agrupamentos de dados já se encontra construído, no CEAC*Boost* cada agrupamento de dados é gerado sequencialmente, considerando o grau de confiança de atribuição de cada objecto aos grupos. O grau de confiança de atribuição de um objecto x_i aos grupos é estimado considerando a similaridade média de x_i ao conjunto dos seus KV vizinhos mais similares, $KVMS_i$, segundo a matriz de co-associações *co_assoc* e uma matriz de normalização F, que contabiliza a frequência com que cada par de objectos é seleccionado, simultaneamente, para a geração de um agrupamento de dados.

Cada agrupamento de dados P^m é obtido usando amostragem de dados com reposição, tendo cada objecto x_i uma probabilidade diferente para ser seleccionado para a amostra, dependendo da confiança de atribuição de x_i aos grupos, nas iterações anteriores. A confiança de atribuição $Conf(x_i)$ de um objecto x_i aos grupos é calculada por

$$Conf(x_i) = \frac{\sum_{j:x_j \in KVMS_i} \frac{co\_assoc_{ij}}{F_{ij}}}{KV} \tag{5.5}$$

em que $Conf(x_i) \in [0, 1]$. Se $Conf(x_i) = 1$, x_i foi agrupado no mesmo grupo que os seus objectos vizinhos mais similares $KVMS_i$, sempre que estes tenham sido seleccionados na mesma amostra de dados. Pretende-se que os objectos mais difíceis de agrupar sejam seleccionados mais vezes para pertencer às amostras de dados, devendo a probabilidade de um objecto x_i ser seleccionado ser inversa à confiança da sua atribuição aos grupos $Conf(x_i)$. Por este motivo, a probabilidade $PS_i \in \mathcal{PS}$ de um objecto x_i ser seleccionado numa amostra de dados é calculada por

$$PS_i = \frac{\exp(1 - Conf(x_i))}{\sum_{j=1}^{N} \exp(1 - Conf(x_j))} \tag{5.6}$$

em que $PS_i \in [0, 1]$ e $\sum_{i=1}^{n} PS_i = 1$.

Para se obter o agrupamento de dados de consenso P^*, é aplicado um algoritmo de agrupamento hierárquico de dados com restrições a *co_assoc*, após esta ter sido normalizada, dividindo cada entrada co_assoc_{ij} por F_{ij}, tal como mostra a equação 5.7.

$$co\_assoc_{ij} = \frac{co\_assoc_{ij}}{F_{ij}} \tag{5.7}$$

O pseudo-código do método de combinação de agrupamentos de dados com restrições CEAC*Boost* é apresentado no algoritmo 5.3.

Algoritmo 5.3: CEAC*Boost*

Entrada: $\mathcal{X} = \{x_1, \cdots, x_n\}$ - Conjunto de n objectos de dados, $Rest_{=}$ - Restrições de ligação obrigatória, $Rest_{\neq}$ - Restrições de ligação proibida, N - Número de agrupamentos de dados a combinar.

Saída: P^* - Agrupamento de dados de consenso

1 /*Inicializações*/;
2 Criar duas matrizes nulas de tamanho $n \times n$, *co_assoc* e F, em que n é o número de objecto ;
3 /*Definir a probabilidade inicial de cada objecto ser seleccionado para uma amostra de dados, $\mathcal{PS} = \{PS_i, \cdots, PS_n\}$*/;
4 $PS_i = \frac{1}{n}$;
5 /*Calcular co-associações entre pares de objectos, com base na distribuição $\mathcal{PS}$*/;
6 **para** $l = 1$ **até** N **faça**
7 Obter um agrupamento de dados P^l aplicando um algoritmo de agrupamento de dados a uma amostra de $\mathcal{X}$, obtida com reposição segundo a distribuição $\mathcal{PS}$;
8 /*Actualizar co-associações entre pares de objectos da amostra colocados no mesmo grupo no agrupamento de dados P^l*/;
9 **para cada** $C_k^l \in P^l$ **faça**
10 **para cada** $(x_i, x_j) \in C_k^l$ **faça**
11 $co\_assoc_{ij} = co\_assoc_{ij} + 1$;
12 **fim**
13 **fim**
14 /*Actualizar frequência com que pares de objectos foram escolhidos, simultaneamente, para a amostra de dados usada para produzir o agrupamento de dados P^l*/;
15 **para cada** $(x_i, x_j) \in P^l$ **faça**
16 $F_{ij} = F_{ij} + 1$;
17 **fim**
18 /*Calcular a confiança de atribuição de cada objecto x_i aos grupos*/;
19 **para cada** $x_i \in \mathcal{X}$ **faça**
20 $Conf(x_i) = \frac{\sum_{j:x_j \in KVMS_i} \frac{co\_assoc_{ij}}{F_{ij}}}{KV}$;
21 **fim**
22 /*Actualizar probabilidade de cada objecto x_i ser escolhido para a o próximo agrupamento de dados P^{l+1}*/;
23 **para cada** $x_i \in \mathcal{X}$ **faça**
24 $PS_i = \frac{\exp(1-Conf(x_i))}{\sum_{j=1}^{N} \exp(1-Conf(x_j))}$;
25 **fim**
26 **fim**
27 /*Normalizar a matriz de co-associações*/;
28 **para cada** $x_i \in \mathcal{X}$ **faça**
29 $co\_assoc_{ij} = \frac{co\_assoc_{ij}}{F_{ij}}$;
30 **fim**
31 Aplicar um algoritmo de agrupamento hierárquico de dados com restrições à matriz de co-associações *co_assoc* para se obter o agrupamento de consenso P^*;

5.4.3 Combinação de Agrupamentos de Dados usando o COP-COBWEB

O terceiro método de combinação de agrupamentos de dados com restrições proposto baseia-se no mapeamento do conjunto de N agrupamentos de dados $\mathcal{P} = \{P^1, \cdots, P^N\}$ numa nova representação do conjunto de dados $\mathcal{X}'$. Cada agrupamento de dados P^i é visto como um meta-atributo categórico MA^i do novo conjunto de dados $\mathcal{X}'$, contendo tantos valores diferentes como o número de grupos em P^i. Assim, a nova representação dos dados $\mathcal{X}' = \{MA^1, \cdots, MA^N\}$ consiste num conjunto de dados com atributos categóricos, pelo que a função de consenso usada para produzir o agrupamento de dados final pode ser qualquer algoritmo de agrupamento de dados categórico que incorpore restrições, como é o caso do COP-COBWEB.

Com esta formulação, pretende-se produzir um agrupamento de dados de consenso P^* que tenha em consideração quer as estruturas dos agrupamentos de dados a combinar, quer as restrições de ligações obrigatória e proibida, sendo as últimas impostas pelo algoritmo de agrupamento de dados com restrições COP-COBWEB.

Algoritmo 5.4: Combinação de Agrupamentos de dados usando o COP-COBWEB

Entrada: $\mathcal{X}$ - Conjunto de n objectos de dados, N - Número de agrupamentos de dados a combinar, $Rest_{=}$ - Restrições de ligação obrigatória, $Rest_{\neq}$ - Restrições de ligação proibida

Saída: P^* - Agrupamento de dados de consenso

1 Criar uma matriz MA de tamanho $n \times N$ para armazenar os meta-atributos da nova representação do conjunto de dados;
2 `/*Obter os meta-atributos da nova representação dos dados usando vários agrupamentos de` $\mathcal{X}$`*/`;
3 **para** $i = 1$ **até** N **faça em paralelo**
4 Produzir um novo agrupamento P^i do conjunto de dados $\mathcal{X}$;
5 Usar os rótulos de cada objecto em P^i como os valores do i-ésimo meta-atributo categórico. $MA^i = P^i$;
6 **fim**
7 Obter o agrupamento de dados de consenso aplicando o algoritmo de agrupamento de dados categóricos com restrições COP-COBWEB à nova representação dos dados $\mathcal{X}'$. $P^* = COP\text{-}COBWEB(\mathcal{X}', Rest_{=}, Rest_{\neq})$

5.4.4 Optimização da Média da Consistência dos Grupos com Penalização de Violações

O último método de agrupamento de dados com restrições proposto baseia-se na optimização de uma função-objectivo $\jmath_{MCGPV}$ baseada na Medida de Consistência dos Grupos [27] (MCG) e na Penalização de Violação (PV) de restrições, usando um algoritmo genético.

A MCG mede a similaridade média entre cada agrupamento de dados a combinar $P^i \in \mathcal{P}$ com o agrupamento de dados de consenso P^*, pressupondo que o número de grupos dos agrupamento de dados a combinar é igual ou superior ao número de grupos K do agrupamento de dados de

consenso $P^* = \{C_1^*, \cdots, C_K^*\}$. A similaridade entre dois agrupamentos de dados, P^i e P^j, em que o número de grupos de P^i é inferior ou igual ao de P^j ($K^i \leq K^j$), é calculada verificando-se se cada um dos K^j grupos pertencentes a $P^j = \{C_1^j, \cdots, C_{K^j}^j\}$ "encaixam" num grupo $C_k^i \in P^i$ pertencente a P^i, tal como ilustra a figura 5.6. As áreas a cinzento representam grupos de um agrupamento de dados P^i enquanto que os objectos numerados representam os grupos de um agrupamento de dados P^j. A figura 5.6 a) mostra um exemplo em que a similaridade entre P^i e P^j é elevada, já que, cada grupo $C_k^j \in P^j$ se encontra incluído num grupo C_m^i. Tal já não acontece na figura 5.6 b), pois os objectos dos grupos de P^j, representados pelos números 1, 5 e 7, encontram-se divididos pelos dois grupos de P^i.

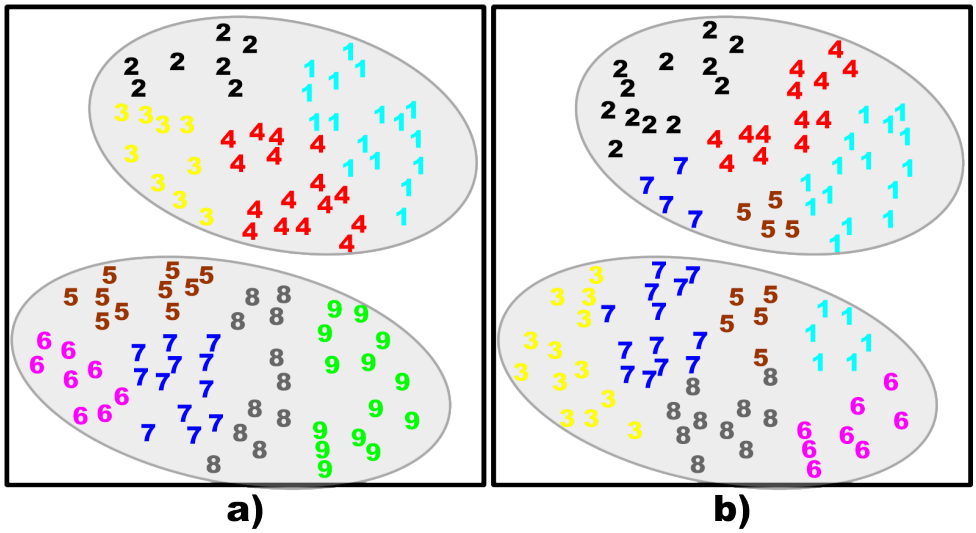

Figura 5.6: Exemplo de similaridade entre dois agrupamentos de dados. - a) Similaridade elevada; b) Similaridade reduzida.

Na MCG, a similaridade entre dois agrupamentos de dados, P^i e P^j, é calculada por

$$simil(P^i, P^j) = \frac{\sum_{l=1}^{K^j} \max_{1 \leq m \leq K^j} (|Inters_{lm}|) \times (1 - \frac{|C_m^i|}{n})}{n}, K^i \leq K^j \tag{5.8}$$

em que $|Inters_{lm}|$ corresponde à cardinalidade do conjunto de objectos comuns ao l-ésimo grupo de P^j e ao m-ésimo grupo de P^i, ou seja, $Inters_{lm} = \{x_a\}, \forall x_a \in C_l^j, x_a \in C_m^i$. Note que na equação 5.8, a intersecção $Inters_{lm}$ entre os dois grupos é ponderada por $(1 - \frac{|C_m^i|}{n})$, com o intuito de evitar que agrupamentos de dados com grupos dominantes (que possuam quase todos os objectos do conjunto de dados) tenham um valor alto de MCG. A Média de Consistência dos Grupos entre um conjunto de agrupamentos de dados $\mathcal{P} = \{P^1, \cdots, P^N\}$ e um agrupamento de consenso P^* é calculada pela equação 5.9.

$$MCG(\mathcal{P}, P^*) = \frac{\sum_{i=1}^{N} simil(P^i, P^*)}{N} \tag{5.9}$$

Para além da MCG, a função-objectivo $\jmath_{MCGPV}$ considera também a fracção de restrições que são satisfeitas pelo agrupamento de consenso, com intuito de penalizar os agrupamentos

de consenso que menos satisfazem os conjuntos de restrições de ligações obrigatória e proibida, $Rest_{=}$ e $Rest_{\neq}$, respectivamente. A penalização de violação de restrições é calculada por

$$PV(P^*, Rest_{=}, Rest_{\neq}) = \frac{\sum_{(x_i,x_j)\in Rest_{=}} I(l_i = l_j) + \sum_{(x_i,x_j)\in Rest_{\neq}} I(l_i \neq l_j)}{|Rest_{=}| + |Rest_{\neq}|} \tag{5.10}$$

em que $|Rest_{=}|$ e $|Rest_{\neq}|$ corresponde ao número de ligações obrigatórias e proibidas, respectivamente, e $I(\cdot)$ toma valor 1 se a expressão for verdadeira e 0 no caso contrário.

Juntando as equações 5.9 e 5.10 numa função-objectivo e ponderando cada uma das medidas, resulta a função-objectivo $\jmath_{MCGPV}$ apresentada na equação 5.11, que se pretende maximizar com o objectivo de encontrar o melhor agrupamento de consenso P^*.

$$\jmath_{MCGPV}(\mathcal{P}, P^*, Rest_{=}, Rest_{\neq}) = \frac{MCG(\mathcal{P}, P^*) + \beta PV(P^*, Rest_{=}, Rest_{\neq})}{1+\beta}, \beta \geq 0 \tag{5.11}$$

A ponderação β da penalização de violação $PV(P^*, Rest_{=}, Rest_{\neq})$ é um parâmetro definido pelo utilizador, tendo como objectivo definir o grau de importância na função-objectivo $\jmath_{MCGPV}$ da consistência entre os grupos e da violação de penalizações. Se $\beta = 1$, $MCG(\mathcal{P}, P^*)$ e $PV(P^*, Rest_{=}, Rest_{\neq})$ têm o mesmo grau de importância na função-objectivo $\jmath_{MCGPV}$ que se pretende maximizar. Se $\beta < 1$, a consistência entre os grupos dos agrupamentos de dados a combinar tem um grau de importância superior à penalização de violações. Se $\beta > 1$, a função-objectivo concede um grau de importância superior à penalização de violações. O denominador $1 + \beta$ é usado para normalizar $\jmath_{MCGPV}$ no intervalo $[0, 1]$.

Para maximizar a função $\jmath_{MCGPV}$, definida na equação 5.11, é utilizado um algoritmo genético. Os algoritmos genéticos são métodos de optimização que se baseiam na evolução dos seres vivos [14]. Estes algoritmos representam possíveis soluções como indivíduos de uma população e optimizam a função-objectivo imitando a evolução da população representada, segundo a *lei do mais forte*. Os indivíduos que melhor se adaptam ao ambiente (os que melhor optimizam a função-objectivo) perduram, enquanto que os restantes desaparecem.

O funcionamento do método de combinação de agrupamentos de dados proposto é explicado de seguida. Inicialmente, é construída a população inicial $\mathcal{B}^0$, isto é, um conjunto de $TamPop$ agrupamentos de dados possíveis, $\mathcal{B}^0 = \{b_1^0, \cdots, b_{TamPop}^0\}$. A pesquisa do agrupamento de dados P^*, que maximiza a função-objectivo $\jmath_{MCGPV}$, é realizada através do cruzamento e mutação da população inicial $\mathcal{B}^0$ e dos indivíduos (agrupamentos de dados) das gerações seguintes. A população inicial pode ser obtida, por exemplo, pela atribuição aleatória de rótulos a cada objecto em cada agrupamento de dados. Outra forma, consiste na execução de algoritmos de agrupamento de dados (com ou sem restrições), com o intuito de dirigir o espaço de pesquisa inicial para um espaço provavelmente mais próximo da melhor solução. Neste trabalho, usou-se o algoritmo de agrupamento de dados K-médias para gerar a população inicial.

Algoritmo 5.5: Média de Consistência dos Grupos com Penalização de Violação

Entrada: $\mathcal{P} = \{P^1, \cdots, P^N\}$ - Conjunto de N agrupamentos de dados, $Rest_=$ - Restrições de ligação obrigatória, $Rest_{\neq}$ - Restrições de ligação proibida, $MaxIter$ - Número máximo de iterações, $TamPop$ - Tamanho da população, $ProbCruz$ - Probabilidade de cruzamento, $ProbMutacao$ - Probabilidade de mutação, $\mathcal{X}$ - Conjunto de dados, K - Número de grupos da população inicial.

Saída: P^* - Agrupamento de dados de consenso

```
/*Obter população inicial, B^0 ;
para i ← 1 até TamPop faça
    b_i^0 ←K-médias(X, K); B^0 ← B^0 ∪ b_i^0;
fim
t ← 0; MaxVal ← −∞; P* ← {} ;
enquanto t < MaxIter faça
    Ordenar b_i^t ∈ B^t por ordem decrescente de Pr_sel(b_j^t);
    /*Seleccionar indivíduos para reprodução*/;
    para i ← 1 até TamPop faça
        val ← Aleatorio[0,1];cont ← 1;
        enquanto cont < TamPop e val < Σ_{j=1}^{cont} Pr_sel(b_j^t) faça
            cont ← cont + 1;
        fim
        b_i^{t+1} ← b_i^t
    fim
    /*Efectuar cruzamento entre indivíduos*/;
    para i ← 1 até TamPop − 1, Passo 2 faça
        se Aleatorio[0,1] < ProbCruz então
            PontoCruza ← Aleatorio[1, n − 1];
            para cont ← PontoCruza + 1 até n faça
                aux ← b_i^{t+1}[cont]; b_i^{t+1}[cont] ← b_{i+1}^{t+1}[cont]; b_{i+1}^{t+1}[cont] ← aux;
            fim
        fim
    fim
    /*Efectuar mutação dos indivíduos*/;
    para i ← 1 até TamPop faça
        para cont ← 1 até n faça
            se Aleatorio[0,1] < ProbMutacao então
                b_i^{t+1}[cont] = Aleatorio{1, ··· , K}
            fim
        fim
    fim
    /*Calcular nova geração de indivíduos*/;
    Seleccionar para B^{t+1} apenas os melhores TamPop indivíduos, segundo
    ȷ_MCGPV(b_i^{t+1}, Rest_=, Rest_≠);
    /*Guardar melhor solução até ao momento*/;
    Seleccionar o individuo com maior valor de ȷ_MCGPV(b_i^{t+1}, Rest_=, Rest_≠),
    P^max ← arg max_{b_i^{t+1} ∈ B^{t+1}} ȷ_MCGPV(b_i^{t+1}) ;
    se ȷ_MCGPV(P^max) > MaxVal então
        P* ← P^max; MaxVal ← ȷ_MCGPV(P^max, Rest_=, Rest_≠) ;
    fim
    t ← t + 1 ;
fim
Devolver P* como o agrupamento de dados de consenso ;
```

Após a população inicial $\mathcal{B}^0$ ter sido construída, o método de combinação de agrupamentos de dados consiste na iteração de quatro passos, até que uma condição (por exemplo, o número máximo de iterações $MaxIter$) de paragem seja satisfeita:

Seleccionar indivíduos para reprodução. Neste passo, são seleccionados $TamPop$ indivíduos b_j^t usando amostragem com reposição de $\mathcal{B}^t$, com a particularidade da probabilidade $Pr_{sel}(b_j^t)$ de cada individuo ser seleccionado não ser uniforme, sendo proporcional ao seu valor da função-objectivo $\jmath_{MCGPV}$. A probabilidade $Pr_{sel}(b_j^t)$ do indivíduo b_j^t ser seleccionado é apresentada na equação 5.12. Note-se que o mesmo individuo pode ser seleccionado mais que uma vez.

$$Pr_{sel}(b_j^t) = \frac{\jmath_{MCGPV}(\mathcal{P}, b_j^t, Rest_{=}, Rest_{\neq})}{\sum_{i=1}^{TamPop} \jmath_{MCGPV}(\mathcal{P}, b_i^t, Rest_{=}, Rest_{\neq})} \tag{5.12}$$

Cruzar indivíduos. Neste passo, os indivíduos seleccionados no passo anterior são vistos como progenitores e são usados para reproduzir novos indivíduos. Os progenitores são agrupados em pares e existe a probabilidade $ProbCruz$ (parâmetro definido pelo utilizador) de haver cruzamento entre os dois. O cruzamento de dois agrupamentos de dados é realizado escolhendo aleatoriamente um ponto de cruzamento $PontoCruz$, isto é, um número inteiro entre 1 e o número de objectos n, e criando dois descendentes da forma ilustrada na figura 5.7. As $PontoCruz$ primeiras posições do vector do primeiro descendente são iguais às primeiras $PontoCruz$ posições do vector do primeiro progenitor e as restantes posições coincidem com as posições $PontoCruz + 1$ a n do segundo progenitor. Já as $PontoCruz$ primeiras posições do vector do segundo descendente são iguais às primeiras $PontoCruz$ posições do vector do segundo progenitor, sendo as restantes posições iguais às posições $PontoCruz + 1$ a n do primeiro progenitor. Note-se que, para se efectuar o cruzamento entre dois indivíduos é inicialmente necessário efectuar a correspodência entre os grupos dos agrupamentos de dados que os indivíduos representam.

Efectuar mutação de indivíduos. Neste passo, o rótulo dos objectos de cada indivíduo (agrupamento de dados) é eventualmente alterado. A probabilidade de uma mutação $ProbMutacao$ acontecer é geralmente muito reduzida, para evitar que a pesquisa no espaço de soluções seja caótica. No entanto, este procedimento é importante porque diminui a probabilidade do método cair em máximos locais.

Calcular nova geração. Finalmente, de todos os indivíduos existentes são escolhidos os $TamPop$ com maior valor de $\jmath_{MCGPV}$, formando este conjunto a população $\mathcal{B}^{t+1}$ da próxima iteração.

O pseudo-código deste método de combinação de agrupamentos de dados é apresentado no algoritmo 5.5.

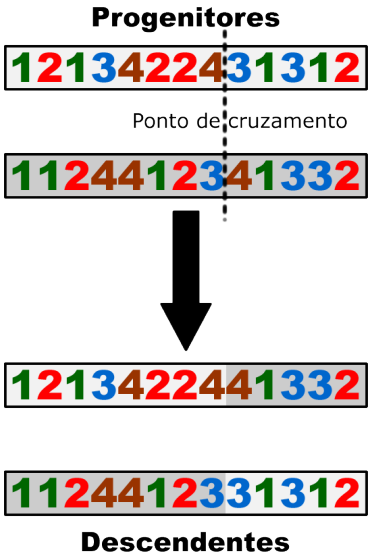

Figura 5.7: Cruzamento de dois agrupamentos de dados - Os dois agrupamentos de dados progenitores são recombinados dando origem a dois descentes. Cada descendente é composto por um subconjunto de rótulos de cada progenitor, sendo esse subconjunto determinado por um ponto de cruzamento.

A figura 5.8 ilustra a evolução do método proposto no conjunto de dados *Iris* (ver secção 6.2 para detalhes), combinando 50 agrupamentos de dados. São usadas no total 50 restrições de ligações obrigatória/proibida, o tamanho da população é $TamPop = 20$, a probabilidade de cruzamento é $ProbCruz = 0.8$ e a probabilidade de mutação é $ProbMutacao = 0.01$. Como se pode verificar, as percentagem de acerto da melhor solução corrente, média de consistência de grupo (MCG) e a penalização de violação (PV) aumentam com o número de iterações, seguindo a tendência da média de $\jmath_{MCGVP}$ da população.

5.5 Sumário

Neste capítulo, foi apresentado a tema da combinação de soluções de algoritmos de aprendizagem, mais precisamente, a combinação de classificadores de dados e a combinação de agrupamentos de dados, sendo descritas vantagens da sua aplicação e as principais abordagens. São propostas: duas versões modificadas do método de Acumulação de Evidências, sendo usados algoritmos de agrupamento de dados hierárquicos com restrições para extrair o agrupamento de consenso da matriz de co-associações; um método que usa o algoritmo de agrupamento de dados categóricos COP-COBWEB, como função de consenso, aplicando o COP-COBWEB a uma nova representação do conjunto de dados, construída usando os rótulos dos objectos em cada agrupamento de dados a combinar; e, finalmente, um método baseado na optimização da Medida de Consistência de Grupos com penalização de violações de restrições usando um algoritmo genético.

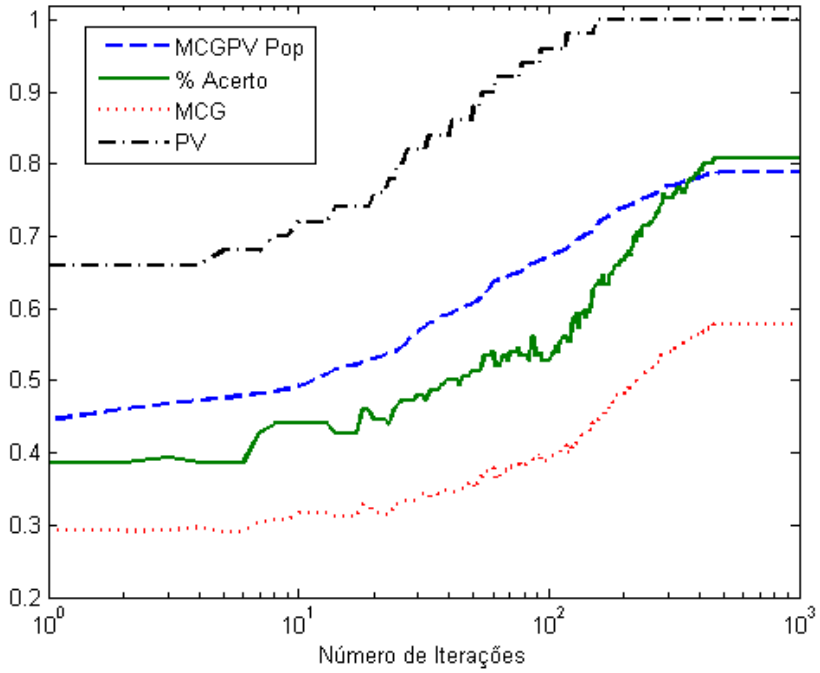

Figura 5.8: Evolução do algoritmo genético para optimizar a função-objectivo $\jmath_{MCGPV}$ - Evolução da $\jmath_{MCGVP}$ média da população (MCGPV pop), percentagem de acerto (% Acerto), média da consistência dos grupos da solução actual (MCG) e penalização de violação da solução actual (PV).

Capítulo 6

Avaliação de Algoritmos de Agrupamento de Dados e Métodos de Combinação

6.1 Introdução

Neste capítulo, são apresentados dois estudos comparativos, com o objectivo de avaliar o desempenho dos algoritmos de agrupamento de dados com restrições e dos métodos de combinação de agrupamentos de dados, comparando os seus resultados com algoritmos de aprendizagem não supervisionada. O primeiro estudo é apresentado na secção 6.5 e abrange oito algoritmos de agrupamento de dados com restrições, previamente descritos no capítulo 4, sendo os desempenhos destes algoritmos comparados com o bem conhecido algoritmo de agrupamento de dados não supervisionado K-médias. No segundo estudo, apresentado na secção 6.6, são avaliados os desempenhos dos métodos de combinação propostos na secção 5.4 do capítulo 5, tendo como referência para comparação o algoritmo de combinação de agrupamentos de dados não supervisionado Acumulação de Evidências (EAC).

Na secção 6.2 são descritos os conjuntos de dados usados nos estudos comparativos acima referidos e na secção 6.3 é apresentada a medida de avaliação dos agrupamentos de dados produzidos pelos algoritmos de agrupamento de dados e pelos métodos de combinação de agrupamentos de dados.

6.2 Conjuntos de Dados

Para se proceder à avaliação dos algoritmos de agrupamento de dados e métodos de combinação de agrupamentos de dados, foram utilizados doze conjuntos de dados bastante distintos, sendo

quatro deles artificiais e os restantes oito reais. Os conjuntos de dados artificiais foram gentilmente cedidos pela Doutora Ana Fred e os conjuntos de dados reais encontram-se num repositório *online* (http://mlearn.ics.uci.edu/MLRepository.html) mantido pela UCI (*University of California - Irvine*).

Com esta variedade de conjuntos de dados, pretende-se avaliar o desempenho dos algoritmos de agrupamento de dados e métodos de combinação de agrupamentos de dados que incorporem restrições numa grande diversidade de situações, tais como, grupos com forma arbitrária, grupos com densidades distintas, grupos bem separados, grupos bastante próximos e conjuntos de dados com diferentes cardinalidades e número de dimensões.

De seguida, apresenta-se a descrição de cada conjunto de dados utilizado nos dois estudos comparativos.

Bars. O conjunto de dados *Bars* é composto por dois grupos bastante próximos, cada um com 200 objectos, com a densidade dos grupos a aumentar da esquerda para a direita, como se pode verificar na figura 6.1.

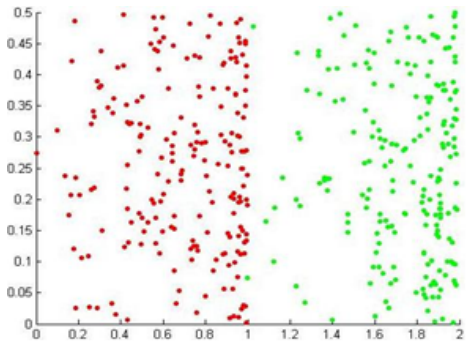

Figura 6.1: Conjunto de dados *Bars* - Composto por 2 grupos, tendo 200 objectos cada grupo.

Cigar. O conjunto de dados *Cigar* é constituído por quatro grupos, possuindo dois dos grupos 100 objectos cada e os restantes dois grupos 25 objectos cada, tal como ilustrado na figura 6.2.

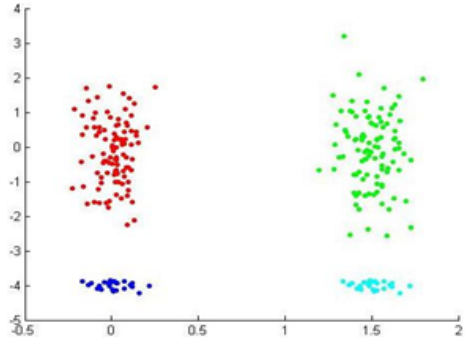

Figura 6.2: Conjunto de dados *Cigar* - Composto por 4 grupos, tendo dois dos grupos 100 objectos e os restantes dois 25 objectos.

Spiral. O conjunto de dados *Spiral* é formado por dois grupos em espiral com 100 objectos cada um. A disposição dos objectos deste conjunto de dados é ilustrada na figura 6.3.

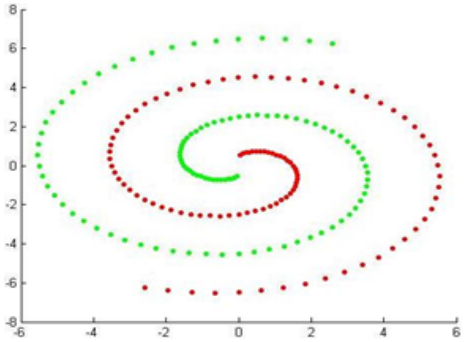

Figura 6.3: Conjunto de dados *Spiral* - Composto por 2 grupos, estando cada grupo disposto em forma de espiral com 100 objectos.

Half Rings. O conjunto de dados *Half Rings* é constituído por três grupos, dois deles com 150 objectos e o terceiro com 200, como se pode verificar na figura 6.4.

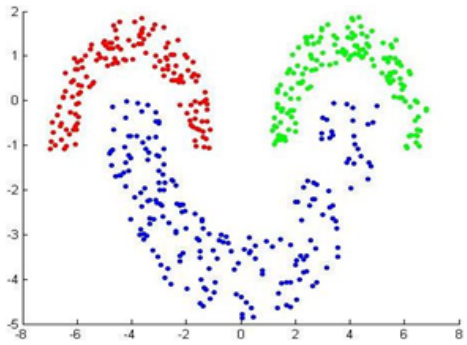

Figura 6.4: Conjunto de dados *Half Rings* - Composto por 3 grupos, tendo dois dos grupos 150 objectos cada um e o terceiro grupo 200 objectos.

Iris. O conjunto de dados *Iris* [4] consiste em três tipos de plantas com 50 objectos de dados para cada tipo. Os dados são caracterizados por quatro atributos, em que, uma das classes se encontra bem separada das outras duas classes que se sobrepõem.

Breast Cancer. O conjunto de dados *Breast Cancer* [90] é composto por 683 objectos, divididos em duas classes, benigno e maligno, e cada objecto é constituído por nove atributos.

Yeast Cell Cycle. O conjunto de dados *Yeast Cell* [63] é constituído por 384 objectos com 17 atributos, divididos em cinco classes que representam cinco fases do ciclo de uma célula. Existem duas versões deste conjunto de dados, a primeira é chamada *Log Yeast* em que é usado o logaritmo do nível de expressão e a outra denominada *Std Yeast*, correspondente a uma versão estandardizada do mesmo conjunto de dados, com média 0 e variância 1.

Optdigits. Consiste num subconjunto do conjunto de dados *Handwritten Digits* [3] (Dígitos Manuscritos). Dos 3823 objectos de dados disponíveis, com 64 atributos, foram usados apenas os primeiros 100 objectos de cada dígito, totalizando 1000 objectos de dados.

Glass. O conjunto de dados *Glass* [39] foi criado para ajudar a investigação criminal, de forma a utilizar o vidro deixado nas cenas de crime como prova. O *Glass* é constituído por 214 objectos de dados, pertencentes a seis classes, caracterizados pela sua composição química e descrita em 9 atributos.

Wine. O conjunto de dados *Wine* [12] é composto por três grupos compostos por 59, 71 e 48 objectos. Os dados correspondem aos resultados de uma análise química de vinhos produzidos na mesma região de Itália, mas são derivados de três cultivações diferentes. Cada objecto é caracterizado por 13 atributos que correspondem às quantidades de 13 constituintes dos vinhos partilhados pelos três tipos de vinho.

Image Segmentation. O conjunto de dados *Image Segmentation* [18] é composto por 2310 objectos de dados, com 19 atributos, em que cada objecto é um segmento de imagem de tamanho 3×3 pixel. Os objectos foram obtidos aleatoriamente a partir de sete imagens ao ar livre e o atributo alvo corresponde a sete superfícies diferentes: parede de tijolo, céu, folhagem, cimento, janela, caminho e relva.

6.3 Medida de Avaliação

Para se avaliar os resultados dos algoritmos de agrupamento de dados e dos métodos de combinação de agrupamentos de dados, os agrupamentos de dados produzidos foram avaliados usando o índice de Consistência [35]. O índice de Consistência compara dois agrupamentos de dados P^1 e P^2, fazendo a correspondência entre os grupos de P^1 e P^2 e medindo em seguida a fracção de objectos partilhados pelos pares de grupos correspondentes nos dois agrupamentos de dados. Formalmente, o índice de Consistência é definido por:

$$iC(P^1, P^2) = \frac{1}{n} \sum_{k=1}^{\min\{K^1, K^2\}} |C_k^1 \cap C_k^2| \tag{6.1}$$

em que $|C_k^i \cap C_k^2|$ corresponde à cardinalidade da intersecção dos objectos dos k-ésimos grupos de P^1 e P^2, assumindo que já foi realizada a correspondência entre os grupos dos dois agrupamentos de dados. Este índice retorna a percentagem de objectos colocados no mesmo grupo, em ambos os agrupamentos de dados P^1 e P^2.

Na avaliação dos resultados dos algoritmos de agrupamentos de dados e dos métodos de combinação de agrupamentos de dados, um dos agrupamentos de dados usado como entrada para o índice de Consistência corresponde à solução encontrada pelo algoritmo de agrupamento de dados ou método de combinação de agrupamentos de dados e o segundo corresponde ao agrupamento "real" do conjunto de dados.

6.4 Geração de Restrições ao Nível dos Objectos de Dados

Para a realização da avaliação dos algoritmos de agrupamentos de dados com restrições foi necessário construír artificialmente vários conjuntos de restrições. Em ambos os estudos apresentados (secções 6.5 a 6.6) são usados conjuntos de 10, 20, 50, 100 e 200 restrições na forma

de relações entre pares de objectos. Para os algoritmos de agrupamento de dados que usam restrições de ligação obrigatória e de ligação proibida foram escolhidos sequencialmente $NumRest \in \{10, 20, 50, 100, 200\}$ pares de objectos de dados (x_i, x_j) com $x_i \neq x_j$, sendo cada par de objectos incluído no conjunto de restrições de ligação obrigatória $Rest_=$ no caso dos rótulos "reais" dos objectos serem iguais e é adicionado ao conjunto de restrições de ligação proibida $Rest_{\neq}$, caso contrário. Para os algoritmos de agrupamento de dados que usam rotulação parcial do conjunto de dados, os rótulos "reais" dos objectos existentes em $Rest_=$ e $Rest_{\neq}$ são usados como entrada.

6.5 Avaliação de Algoritmos de Agrupamento de Dados

Nesta secção são apresentados os resultados dos algoritmos de agrupamento de dados com restrições e do algoritmo K-médias para os 12 conjuntos de dados descritos na secção 6.2. Na subsecção 6.5.1 é descrita a configuração experimental desta avaliação, sendo os resultados para cada conjunto de dados apresentados nas subsecções 6.5.2 a 6.5.13. Na subsecção 6.5.14 é efectuado um breve resumo dos resultados apresentados.

6.5.1 Configuração Experimental

Nesta avaliação são comparados os algoritmos de agrupamento de dados com restrições Cop-K-médias, K-médias Semeado (K-médias Sem), K-médias Restringido (K-médias Rest), PCK-médias, Ligação Completa Restringido (CCL - *Constrained Complete Link*), *Constrained Vector Quantization Error* (CVQE), *Linear-time Constrained Vector Quantization Error* (LCVQE) e o MPCK-médias, e o algoritmo de agrupamento de dados não supervisionado K-médias. Para cada conjunto de dados, são gerados conjuntos de restrições tal como descrito na secção 6.4. Foram usados como entrada para os algoritmos de agrupamento de dados apenas os conjuntos de dados, os respectivos números "reais" de grupos e os conjuntos de restrições. Todas as experiências foram repetidas 20 vezes, sendo gerados novos conjuntos de restrições em cada experiência. Para cada conjunto de dados, são apresentados os melhores resultados por algoritmo de agrupamento de dados e respectivos resultados médios. Realça-se o facto de que no algoritmo de agrupamento de dados com restrições MPCK-médias, a aprendizagem da medida de distância foi realizada considerando apenas a diagonal da matriz de parametrização. Nos conjuntos de dados *Optdigits* e *Image Segmentation* não são apresentados resultados para o algoritmo MPCK-médias, porque não foi possível obter em tempo útil os respectivos resultados, devido ao custo computacional desse algoritmo e às elevadas cardinalidades e dimensionalidades dos conjuntos de dados referidos.

6.5.2 Resultados dos Algoritmos de Agrupamento para o Conjunto de Dados *Bars*

Método	10	20	50	100	200
Cop-K-médias	97.25%	95.25%	96.75%	-	-
K-médias Sem	98.5%	98%	98.5%	98%	98.5%
K-médias Rest	**98.75%**	98%	98.5%	**98.75%**	**99.75%**
PCK-médias	**98.75%**	98.5%	98.5%	**98.75%**	**99.75%**
CCL	**98.75%**	**98.75%**	**99.5%**	69.5%	65%
CVQE	**98.75%**	98.5%	98.5%	**98.75%**	**99.75%**
LCVQE	98.5%	98.5%	98.5%	98.5%	98.75%
MPCK-médias	98.25%	98%	98%	98.5%	**99.75%**
K-médias	98.5%	98.5%	98.5%	98.5%	98.5%

Tabela 6.1: Valores máximos de iC para os algoritmos de agrupamento no conjunto de dados *Bars*.

Na tabela 6.1 são apresentados os melhores resultados obtidos pelo algoritmo K-médias e pelos algoritmos de agrupamento de dados com restrições para o conjunto de dados *Bars*. Os algoritmos obtiveram resultados máximos semelhantes, com valores de índice de Consistência (iC) a variar entre 95% e 99.75%. O K-médias obteve como valor máximo 98.5%, um valor um pouco inferior aos obtidos pelos algoritmos de agrupamento de dados com restrições K-médias Restringido, PCK-médias, CVQE e MPCK-médias, 99.75%. O único algoritmo de agrupamento de dados com restrições que não obteve valores acima dos 95% foi o CCL com 100 e 200 restrições, alcançando apenas 69.5% e 65%, respectivamente. Note-se que o Cop-K-médias não conseguiu realizar o agrupamento de dados com 100 e 200 restrições, facto assinalado na tabela com o sinal -. A figura 6.5 apresenta os resultados médios dos algoritmos acima referidos neste conjunto de dados. Como se pode verificar, os resultados médios são idênticos aos resultados máximos. Os valores 0 indicam que o Cop-K-médias não conseguiu gerar quaisquer agrupamentos.

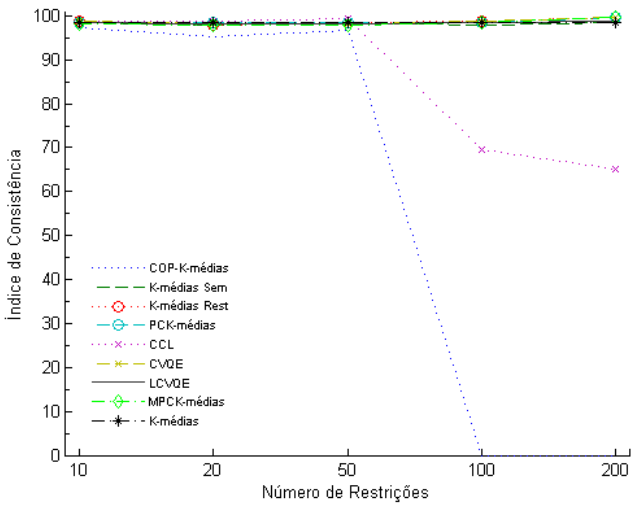

Figura 6.5: Resultados dos algoritmos de agrupamento para o conjunto de dados *Bars* - Valores médios do índice de Consistência.

6.5.3 Resultados dos Algoritmos de Agrupamento para o Conjunto de Dados *Cigar*

Método	10	20	50	100	200
Cop-K-médias	-	68%	95.2%	-	-
K-médias Sem	**97.6%**	87.6%	**97.6%**	97.6%	97.6%
K-médias Rest	**97.6%**	**98%**	**97.6%**	**99.6%**	**99.6%**
PCK-médias	**97.6%**	65.6%	**97.6%**	69.2%	76.4%
CCL	51.2%	57.6%	62.8%	99.2%	99.2%
CVQE	**97.6%**	97.6%	68.8%	98.4%	69.2%
LCVQE	**97.6%**	68.8%	**97.6%**	98.4%	98%
MPCK-médias	**97.6%**	70%	**97.6%**	97.6%	98.8%
K-médias	63.35%	63.35%	63.35%	63.35%	63.35%

Tabela 6.2: Valores máximos de iC para os algoritmos de agrupamento no conjunto de dados *Cigar*.

No conjunto de dados *Cigar*, os melhores resultados dos algoritmos de agrupamento de dados que usam restrições foram claramente superiores ao melhor resultado do K-médias (tabela 6.2). O algoritmo K-médias obteve como melhor valor para este conjunto de dados 63.25% de iC, enquanto que os algoritmos que usam restrições obtiveram em pelo menos uma situação valores superiores a 95% de iC. Neste conjunto de dados, destaca-se o algoritmo K-médias Restringido, que obteve sempre os melhores resultados, independentemente do número de restrições, tendo obtido 99.6% como melhor resultado, usando tanto 100 como 200 restrições. Relativamente aos resultados médios, apresentados na figura 6.6, destacam-se os algoritmos K-médias Semeado e K-médias Restringido, seja qual for o número de restrições, e o CCL quando o número de restrições é igual ou superior a 100, pois obtêm valores médios de iC geralmente superiores a 95%.

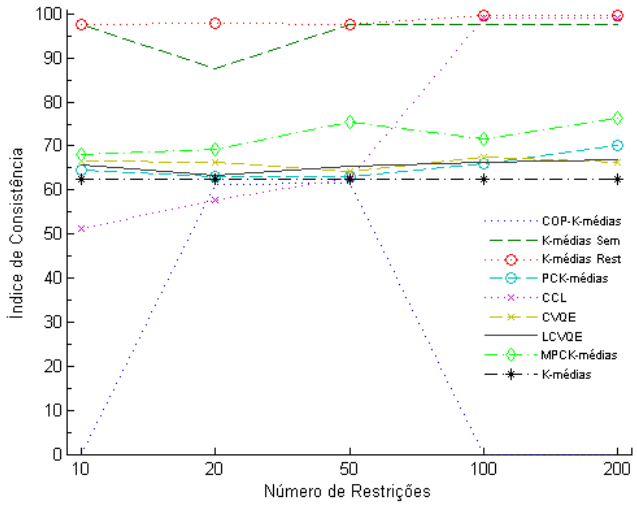

Figura 6.6: Resultados dos algoritmos de agrupamento para o conjunto de dados *Cigar* - Valores médios do índice de Consistência.

6.5.4 Resultados dos Algoritmos de Agrupamento para o Conjunto de Dados *Spiral*

Método	10	20	50	100	200
Cop-K-médias	60.5%	61%	-	-	-
K-médias Sem	55%	57.5%	57.5%	58%	57.5%
K-médias Rest	61.5%	**67**%	62%	86%	93%
PCK-médias	58.5%	62.5%	61.5%	71.5%	87%
CCL	51%	63%	54%	**86.5**%	**99.5**%
CVQE	**62**%	61%	62.5%	60%	62.5%
LCVQE	60.5%	60%	62%	58.5%	58%
MPCK-médias	56%	56.5%	**63.5**%	69.5%	92.5%
K-médias	56.5%	56.5%	56.5%	56.5%	56.5%

Tabela 6.3: Valores máximos de iC para os algoritmos de agrupamento no conjunto de dados *Spiral*.

Os melhores resultados obtidos para o conjunto de dados *Spiral* são apresentados na tabela 6.3. Na globalidade, o desempenho dos algoritmos de agrupamento de dados, com e sem restrições, é bastante mau, variando geralmente entre os 55% e os 70% de iC. As excepções são os algoritmos CCL com 100 e 200 restrições, alcançando 86.5% e 99.5%, respectivamente, PCK-médias com 71.5% e 87% de iC, também com 100 e 200 restrições, e os K-médias Restringido e MPCK-médias com 93% e 92.5%, respectivamente, com 200 restrições. O resultado do K-médias foi de apenas 56.5% de iC. A figura 6.7 apresenta os resultados médios obtidos pelos mesmos algoritmos. Os algoritmos K-médias Restringido, CCL, PCK-médias e MPCK-médias apresentam tendências crescentes de qualidade com o aumento do número de restrições, o que permite concluir que um número elevado de relações entre objectos facilita a descoberta de grupos com formas arbitrárias.

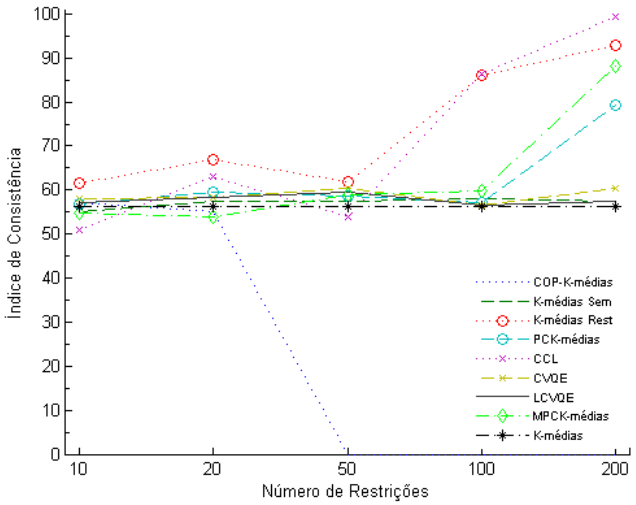

Figura 6.7: Resultados dos algoritmos de agrupamento para o conjunto de dados *Spiral* - Valores médios do índice de Consistência.

6.5.5 Resultados dos Algoritmos de Agrupamento para o Conjunto de Dados *Half Rings*

Método	10	20	50	100	200
Cop-K-médias	77.6%	-	77%	-	-
K-médias Sem	78%	78%	78%	78.4%	78.4%
K-médias Rest	78.4%	79.8%	78.6%	85.4%	88.2%
PCK-médias	78.6%	77%	78.4%	78.2%	80%
CCL	84.2%	**87.8**%	**87.6**%	**98.4**%	75%
CVQE	78.6%	78.6%	78.6%	77.2%	80.2%
LCVQE	78.6%	78.4%	78.6%	78.4%	79.6%
MPCK-médias	**87.2**%	80.4%	87%	88.6%	**90.8**%
K-médias	77.4%	77.4%	77.4%	77.4%	77.4%

Tabela 6.4: Valores máximos de iC para os algoritmos de agrupamento no conjunto de dados *Half Rings*.

No conjunto de dados *Half Rings*, os melhores resultados obtidos pelos algoritmos de agrupamento de dados com restrições e com o algoritmo K-médias variam entre 77% *de* iC no PCK-médias e 98.4% no CCL, sendo este último o melhor valor absoluto obtido. É possível verificar que os algoritmos de agrupamento de dados com restrições obtêm quase sempre melhores resultados que o K-médias, apesar da melhoria não ser geralmente muito significativa, com as excepções dos algoritmos CCL, MPCK-médias e o K-médias Restringido (com 100 e 200 restrições). Os resultados médios obtidos por todos os algoritmos acima referidos para o conjunto de dados *Half Rings* são ilustrados na figura 6.8. Como se pode verificar, os resultados dos algoritmos de agrupamento de dados com restrições são normalmente idênticos ou superiores aos resultados do K-médias, com a excepção do Cop-K-Médias que apenas conseguiu produzir agrupamentos do conjunto de dados com 10 e 50 restrições.

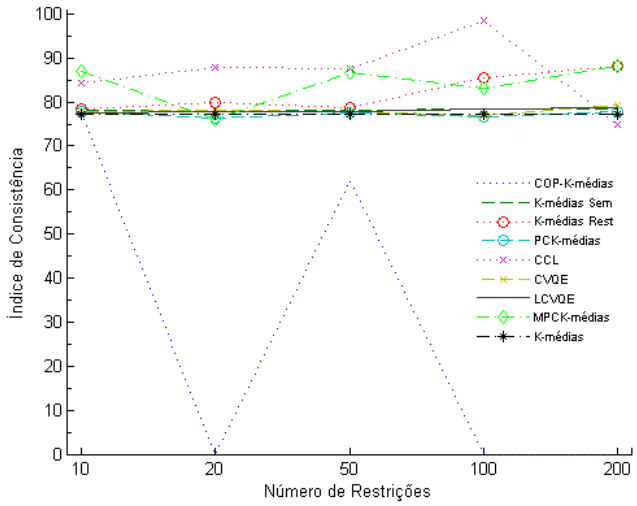

Figura 6.8: Resultados dos algoritmos de agrupamento para o conjunto de dados *Half Rings* - Valores médios do índice de Consistência.

6.5.6 Resultados dos Algoritmos de Agrupamento para o Conjunto de Dados *Iris*

Método	10	20	50	100	200
Cop-K-médias	85.33%	81.33%	86%	-	-
K-médias Sem	88.67%	88.67%	88.67%	88.67%	88.67%
K-médias Rest	92.67%	93.33%	90.67%	97.33%	**100%**
PCK-médias	90%	90%	90%	97.33%	99.33%
CCL	96.67%	78.67%	89.33%	93.33%	99.33%
CVQE	90%	93.33%	90%	93.33%	92.67%
LCVQE	89.33%	92%	89.33%	89.33%	90%
MPCK-médias	**98%**	**99.33%**	**97.33%**	**100%**	99.33%
K-médias	82.47%	82.47%	82.47%	82.47%	82.47%

Tabela 6.5: Valores máximos de iC para os algoritmos de agrupamento no conjunto de dados *Iris*.

A tabela 6.5 apresenta os melhores resultados obtidos pelo algoritmo K-médias e pelos algoritmos de agrupamento de dados com restrições no conjunto de dados *Iris*. Os algoritmos de agrupamento de dados que incorporam restrições têm sistematicamente resultados superiores ao melhor resultado obtido pelo K-médias, que atingiu apenas 82.47% de iC. Neste conjunto de dados realça-se o desempenho do MPCK-médias, que obteve 98%, 99.33%, 97.33% e 100% de iC com 10, 20, 50 e 100 restrições respectivamente, e do K-médias Restringido, que alcançou 100% usando 200 restrições. Na figura 6.9 são ilustrados os resultados médios obtidos pelos mesmos algoritmos neste conjunto de dados. Mais uma vez, o resultado médio de iC obtido pelo algoritmo K-médias é geralmente inferior aos resultados médios dos algoritmos de agrupamento de dados com restrições. As excepções são o algoritmo Cop-K-médias, que não conseguiu produzir agrupamentos deste conjunto de dados com 100 ou mais restrições, e o algoritmo LCVQE.

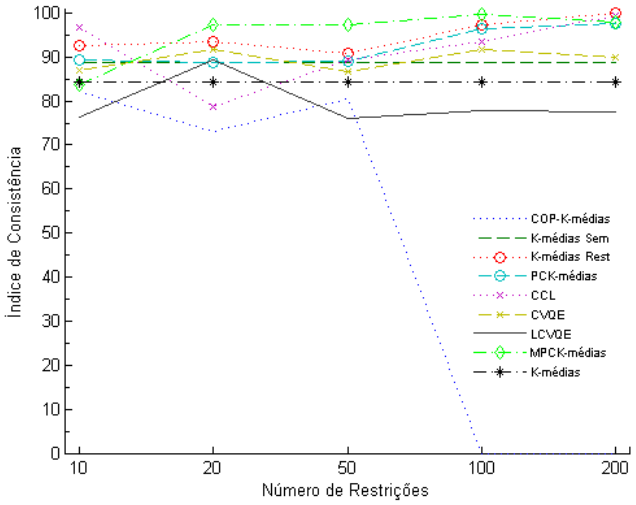

Figura 6.9: Resultados dos algoritmos de agrupamento para o conjunto de dados *Iris* - Valores médios do índice de Consistência.

6.5.7 Resultados dos Algoritmos de Agrupamento para o Conjunto de Dados *Breast Cancer*

Método	10	20	50	100	200
Cop-K-médias	96.05%	-	95.46%	-	-
K-médias Sem	96.19%	96.19%	96.05%	96.19%	96.19%
K-médias Rest	96.34%	**96.49%**	96.05%	**97.07%**	**97.66%**
PCK-médias	96.34%	**96.49%**	**96.19%**	96.78%	97.51%
CCL	**96.63%**	69.99%	92.83%	92.68%	67.35%
CVQE	96.34%	**96.49%**	**96.19%**	**97.07%**	97.36%
LCVQE	96.05%	96.19%	**96.19%**	96.34%	96.78%
MPCK-médias	95.46%	95.61%	69.25%	96.34%	96.49%
K-médias	96.05%	96.05%	96.05%	96.05%	96.05%

Tabela 6.6: Valores máximos de iC para os algoritmos de agrupamento no conjunto de dados *Breast Cancer*.

Na tabela 6.6 são apresentadas os melhores resultados obtidos pelos algoritmos de agrupamento de dados no conjunto de dados *Breast Cancer*. Os resultados são bastante bons em todos os algoritmos de agrupamento de dados com restrições e no K-médias, variando entre os 95.46% e os 97.66%, existindo apenas duas excepções: o CCL com 20 restrições (69.99%) e o MPCK-médias com 50 restrições (69.25%). Apesar de não existir uma diferença evidente entre os resultados do K-médias e dos algoritmos de agrupamento de dados com restrições, realça-se o facto de que os melhores valores obtidos usando conjuntos de restrições com 10, 20, 50, 100 e 200 restrições foram alcançados só por algoritmos de agrupamento de dados que incorporam restrições. No entanto, considerando os valores médios de iC, apresentados na figura 6.10, verifica-se que os resultados do MPCK-médias e do CVQE são inferiores ao resultados obtidos pelo K-médias, indicando que a utilização inadequada de restrições pode ser prejudicial.

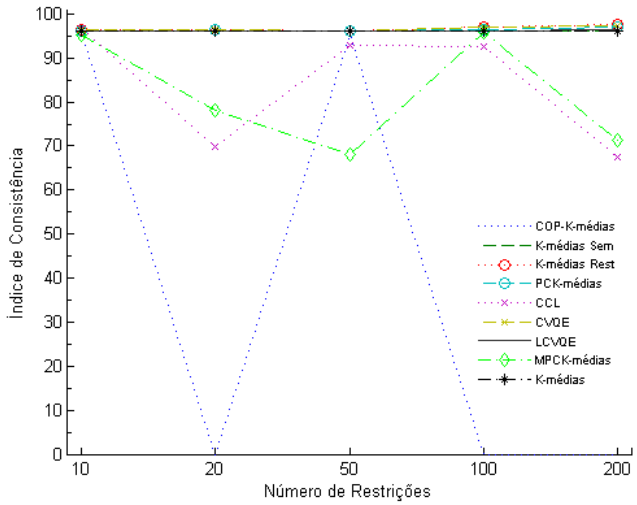

Figura 6.10: Resultados dos algoritmos de agrupamento para o conjunto de dados *Breast Cancer* - Valores médios do índice de Consistência.

6.5.8 Resultados dos Algoritmos de Agrupamento para o Conjunto de Dados *Log Yeast*

Método	10	20	50	100	200
Cop-K-médias	-	35.16%	35.68%	-	-
K-médias Sem	30.73%	32.03%	30.73%	30.73%	33.33%
K-médias Rest	32.55%	35.42%	31.51%	**64.06**%	**77.6**%
PCK-médias	29.17%	29.43%	32.55%	33.59%	31.77%
CCL	28.65%	33.59%	29.69%	32.55%	34.64%
CVQE	**37.5**%	33.33%	**35.94**%	36.2%	34.11%
LCVQE	32.81%	34.38%	35.16%	35.68%	35.16%
MPCK-médias	34.64%	**38.8**%	34.38%	44.27%	40.36%
K-médias	30%	30%	30%	30%	30%

Tabela 6.7: Valores máximos de iC para os algoritmos de agrupamento no conjunto de dados *Log Yeast*.

Os resultados máximos dos algoritmos de agrupamento de dados que incorporam restrições e do algoritmo K-médias são de pouca qualidade, como mostra a tabela 6.7. O melhor desempenho do K-médias obteve apenas 30% de iC, sendo os resultados dos algoritmos de agrupamento de dados com restrições um pouco superiores na sua generalidade. Os agrupamentos com melhor qualidade foram obtidos pelo K-médias Restringido com 64.06% e 77.6% usando conjuntos de restrições com 100 e 200 restrições, respectivamente. Os resultados médios dos algoritmos de agrupamento de dados alcançaram valores de iC entre os 25% e os 35%, excepto os resultados médios do K-médias Restringido, com 100 e 200 restrições, que obteve valores de iC médios superiores a 60% e 75%, respectivamente.

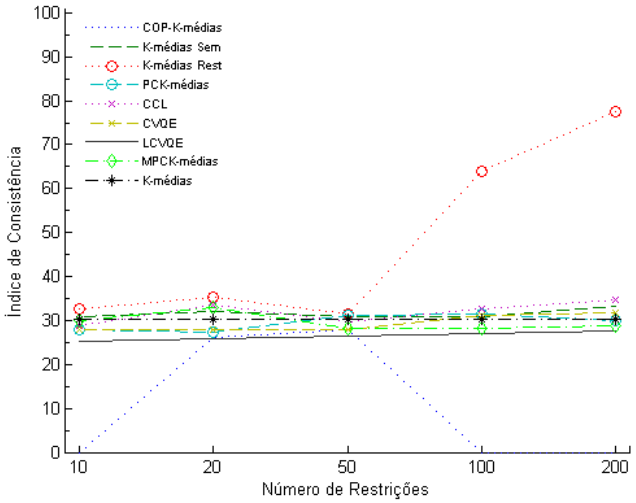

Figura 6.11: Resultados dos algoritmos de agrupamento para o conjunto de dados *Log Yeast* - Valores médios do índice de Consistência.

6.5.9 Resultados dos Algoritmos de Agrupamento para o Conjunto de Dados *Std Yeast*

Método	10	20	50	100	200
Cop-K-médias	72.66%	-	-	-	-
K-médias Sem	65.89%	65.63%	72.92%	73.44%	72.92%
K-médias Rest	65.63%	69.01%	**80.47%**	**86.46%**	**91.93%**
PCK-médias	73.7%	56.25%	56.25%	73.96%	61.46%
CCL	65.63%	66.93%	46.09%	37.76%	36.2%
CVQE	**73.96%**	**75.26%**	75.52%	77.08%	76.3%
LCVQE	73.18%	73.7%	72.92%	73.7%	76.04%
MPCK-médias	33.33%	38.54%	33.33%	35.42%	39.32%
K-médias	63.19%	63.19%	63.19%	63.19%	63.19%

Tabela 6.8: Valores máximos de iC para os algoritmos de agrupamento no conjunto de dados *Std Yeast*.

No conjunto de dados *Std Yeast*, o conjunto de dados original *Yeast Cell Cycle* não é logaritmizado mas sim estandardizado. Esta diferença no pré-processamento dos dados teve um impacto muito significativo no desempenho em todos os algoritmos, como se pode verificar na tabela 6.8. O K-médias alcançou apenas 63.19% de iC, valor claramente ultrapassado pelo desempenho do K-médias Restringido que obteve 91.93% usando um conjunto de 200 restrições. De realçar o péssimo desempenho do algoritmo MPCK-médias que nunca atingiu valores de iC superiores a 39.32%. Na figura 6.12 estão ilustrados os resultados médios obtidos pelos algoritmos de agrupamento de dados com restrições e pelo K-médias. Neste conjunto de dados, o K-médias Semeado e o K-médias Restringido melhoram os seus desempenhos com o aumento do número de restrições. No entanto, outros algoritmos mostram tendências contrárias, existindo também algoritmos sem qualquer tendência.

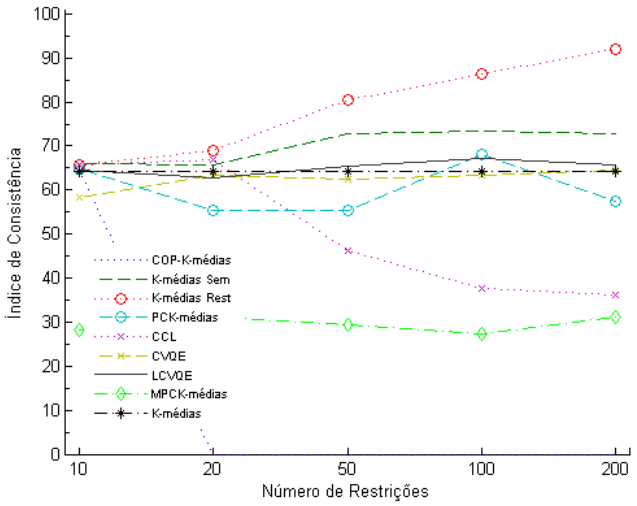

Figura 6.12: Resultados dos algoritmos de agrupamento para o conjunto de dados *Std Yeast* - Valores médios do índice de Consistência.

6.5.10 Resultados dos Algoritmos de Agrupamento para o Conjunto de Dados *Optdigits*

Método	10	20	50	100	200
Cop-K-médias	**81.7%**	79.6%	-	-	-
K-médias Sem	75.6%	82.2%	86.6%	86.5%	86.6%
K-médias Rest	80.4%	**82.6%**	**88.5%**	**92.8%**	**94.8%**
PCK-médias	77.2%	68.8%	69.2%	84.2%	74.8%
CCL	51.8%	65.4%	57.9%	65.6%	68.3%
CVQE	81%	77.7%	78.9%	80.3%	80.4%
LCVQE	74.1%	77.4%	79.6%	74.3%	81.3%
K-médias	73.9%	73.9%	73.9%	73.9%	73.9%

Tabela 6.9: Valores máximos de iC para os algoritmos de agrupamento no conjunto de dados *Optdigits*.

A tabela 6.9 apresenta os melhores resultados obtidos com os algoritmos de agrupamento de dados que incorporam restrições e com o algoritmo K-médias no conjunto de dados *Optdigits*. Relativamente aos valores máximos de iC em cada algoritmo, todos os algoritmos de agrupamento de dados com restrições, com a excepção do CCL, alcançam resultados superiores ao resultado do K-médias, que obteve apenas 73.9% de iC. Neste conjunto de dados destaca-se o desempenho do K-médias Restringido que obteve os melhores resultados usando conjuntos de restrições com 20 (82.6%), 50 (88.5%), 100 (92.8%) e 200 (94.8%). No entanto, considerando os valores médios obtidos pelos algoritmos de agrupamento de dados, apenas os algoritmos PCK-médias e K-médias Restringido alcançaram sempre valores de iC superiores aos obtidos pelo K-médias, como se pode verificar na figura 6.13. No caso deste último algoritmo, consegue-se identificar uma tendência clara no aumento da qualidade do agrupamento de dados com o aumento do número de restrições.

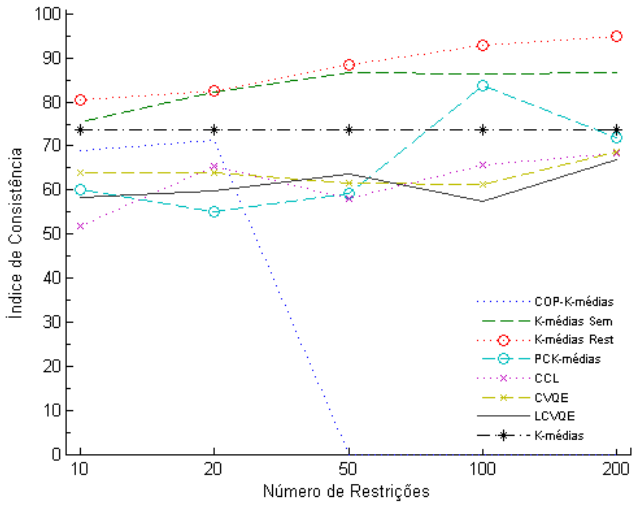

Figura 6.13: Resultados dos algoritmos de agrupamento para o conjunto de dados *Optdigits* - Valores médios do índice de Consistência.

6.5.11 Resultados dos Algoritmos de Agrupamento para o Conjunto de Dados *Glass*

Método	10	20	50	100	200
Cop-K-médias	53.27%	48.6%	-	-	-
K-médias Sem	53.74%	54.21%	55.14%	49.07%	50.93%
K-médias Rest	49.53%	54.21%	**60.75%**	**75.23%**	**92.06%**
PCK-médias	55.61%	**54.67%**	55.61%	53.74%	63.08%
CCL	52.8%	44.86%	39.72%	37.38%	47.66%
CVQE	54.67%	53.74%	54.67%	55.14%	54.21%
LCVQE	54.67%	54.21%	52.8%	53.74%	54.21%
MPCK-médias	**57.3%**	49.53%	46.73%	48.13%	61.21%
K-médias	51.78%	51.78%	51.78%	51.78%	51.78%

Tabela 6.10: Valores máximos de iC para os algoritmos de agrupamento no conjunto de dados *Glass*.

Na tabela 6.10 são apresentados os resultados para o conjunto de dados *Glass*. Neste conjunto de dados, o desempenho de todos os algoritmos de agrupamento de dados é de pouca qualidade. O K-médias obteve 51.78% de iC como melhor resultado, resultado que foi sempre superado pelos algoritmos de agrupamento de dados com restrições, com as excepções do MPCK-médias usando 100 restrições ou menos, do CCL usando conjuntos com 20 restrições ou mais e o COP-K-médias com conjuntos de 20 restrições. A figura 6.14 apresenta os resultados médios obtidos pelos algoritmos de agrupamentos de dados no conjunto de dados *Glass*. Apenas o algoritmo de agrupamento de dados com restrições K-médias Restringido obtém sempre resultados superiores ao K-médias, evidenciando uma vez mais uma tendência crescente de qualidade com o aumento do número de restrições.

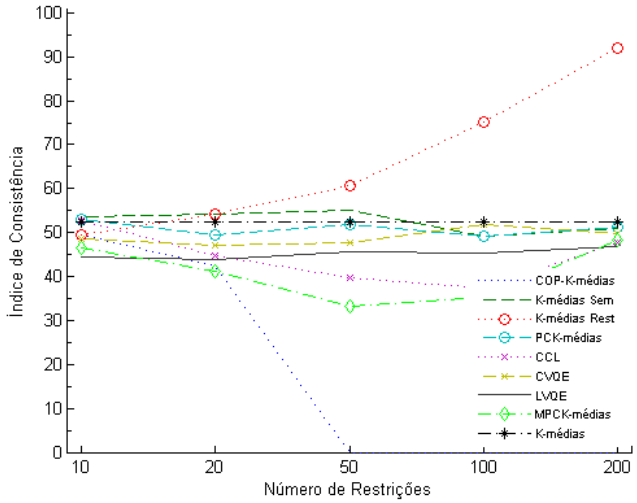

Figura 6.14: Resultados dos algoritmos de agrupamento para o conjunto de dados *Glass* - Valores médios do índice de Consistência.

6.5.12 Resultados dos Algoritmos de Agrupamento para o Conjunto de Dados *Wine*

Método	10	20	50	100	200
Cop-K-médias	66.85%	-	-	-	-
K-médias Sem	70.22%	70.22%	70.22%	70.22%	70.22%
K-médias Rest	72.47%	76.97%	79.78%	92.7%	94.38%
PCK-médias	71.35%	72.47%	73.03%	76.97%	86.52%
CCL	60.67%	69.1%	47.19%	43.82%	65.17%
CVQE	70.22%	70.79%	69.1%	76.4%	75.84%
LCVQE	70.22%	70.22%	69.66%	68.54%	70.22%
MPCK-médias	**96.07**	**96.07%**	**100%**	**100%**	**100%**
K-médias	66.75%	66.75%	66.75%	66.75%	66.75%

Tabela 6.11: Valores máximos de iC para os algoritmos de agrupamento no conjunto de dados *Wine*.

No conjunto de dados *Wine*, os melhores resultados obtidos por todos os algoritmos de agrupamento de dados com restrições superaram sempre o melhor desempenho do K-médias (66.75%), com a excepção do CCL, como mostra a tabela 6.11. Neste conjunto de dados destaca-se o desempenho do MPCK-médias que obteve 100% de iC com conjuntos de 50, 100 e 200 restrições, o que indica que a aprendizagem da medida de distância foi bastante útil para agrupar este conjunto de dados. A figura 6.15 ilustra os resultados médios obtidos pelos algoritmos de agrupamento de dados com restrições e pelo K-médias. Como se pode verificar, os algoritmos Cop-K-Médias, CVQE e LCVQE obtêm regularmente resultados médios inferiores ao K-médias, tendo os algoritmos K-médias Semeados e, principalmente, o MPCK-médias, desempenhos claramente superiores ao K-médias.

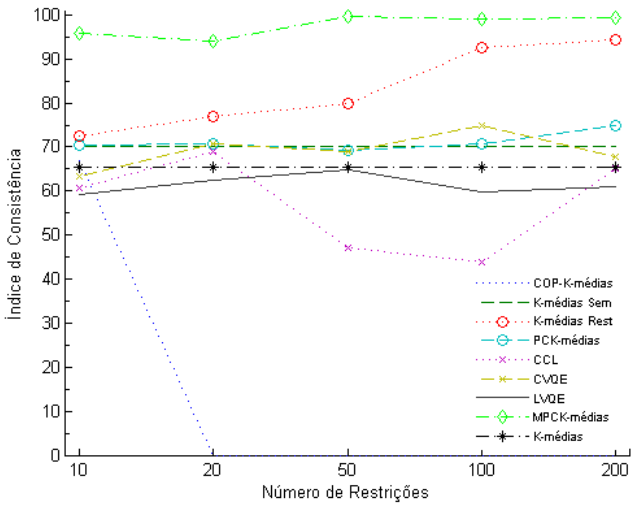

Figura 6.15: Resultados dos algoritmos de agrupamento para o conjunto de dados *Wine* - Valores médios do índice de Consistência.

6.5.13 Resultados dos Algoritmos de Agrupamento para o Conjunto de Dados *Image Segmentation*

Método	10	20	50	100	200
Cop-K-médias	29%	28.66%	28.79%	14.72%	14.68%
K-médias Sem	58.01%	55.67%	63.16%	63.12%	63.38%
K-médias Rest	**58.23%**	**57.01%**	**63.98%**	**66.23%**	**69.7%**
PCK-médias	35.67%	50.09%	57.06%	55.97%	56.1%
CCL	34.94%	35.37%	48.18%	54.85%	47.62%
CVQE	29.09%	46.49%	29.05%	35.45%	35.54%
LCVQE	35.67%	35.63%	35.63%	35.63%	41.04%
K-médias	52.70%	52.70%	52.70%	52.70%	52.70%

Tabela 6.12: Valores máximos de iC para os algoritmos de agrupamento no conjunto de dados *Image Segmentation*.

Os melhores resultados obtidos para o conjunto de dados *Image Segmentation*, o último conjunto de dados apresentado neste estudo comparativo, são apresentados na tabela 6.12. Os resultados são em geral de pouca qualidade, variando entre os 14.68% no algoritmo de agrupamento de dados Cop-K-médias e os 69.7% de iC no K-médias Restringido, que obteve o agrupamento do conjunto de dados com melhor qualidade. Neste conjunto de dados, pode-se concluir que a utilização de restrições no agrupamento de dados nem sempre tem resultados positivos, já que, os algoritmos de agrupamento com restrições Cop-K-médias, CCL, CVQE e LCVQE obtêm normalmente desempenhos inferiores ao K-médias. Na figura 6.16 são ilustrados os resultados médios obtidos para este conjunto de dados. Como mostra a figura, os desempenhos dos algoritmos de agrupamento de dados com restrições são inferiores ao desempenho do K-médias, exceptuando os algoritmos K-médias Semeado e K-médias Restringido.

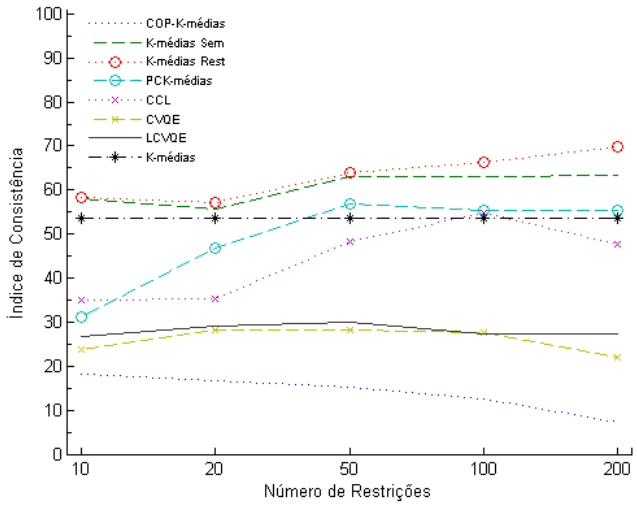

Figura 6.16: Resultados dos algoritmos de agrupamento para o conjunto de dados *Image Segmentation* - Valores médios do índice de Consistência.

6.5.14 Resumo dos Resultados dos Algoritmos de Agrupamento de Dados

Com a análise dos resultados dos algoritmos de agrupamento de dados com restrições, tendo como referência o K-médias como algoritmo de agrupamento de dados não supervisionado, pode-se concluir que geralmente os algoritmos de agrupamento de dados que usam restrições obtêm melhores resultados que os algoritmos não supervisionados, neste caso, que o K-médias. No entanto, o uso de restrições nem sempre contribui para o aumento da qualidade dos agrupamentos de dados, como ficou patente nos conjuntos de dados *Log Yeast*, *Std Yeast*, *Glass* e *Image Segmentation*. O algoritmo com melhor desempenho global foi o algoritmo de agrupamento de dados com restrições K-médias Restringido, que obteve em 10 dos 12 conjuntos de dados usados neste estudo os melhores resultados absolutos segundo o índice de Consistência, não tendo alcançado os melhores resultados apenas nos conjuntos de dados *Half Rings* e *Spiral*. Nestes conjuntos de dados, o algoritmo CCL obteve os melhores desempenhos. Pode-se também concluir que o uso de sementes para inicializar os centros dos grupos no algoritmo K-médias é bastante proveitoso, já que, os algoritmos de agrupamento de dados com restrições K-médias Semeado e K-médias Restringido obtêm resultados quase sempre superiores aos resultados do K-médias, independentemente do número de restrições, e em todos os conjuntos de dados usados neste estudo.

6.6 Avaliação de Métodos de Combinação de Agrupamentos de Dados

Nesta secção são apresentados os resultados obtidos com os métodos de combinação de agrupamentos de dados, com e sem restrições, para os mesmos conjuntos de dados usados na avaliação anterior. Na subsecção 6.6.1 é descrita a configuração experimental desta avaliação, sendo os resultados para cada conjunto de dados apresentados nas subsecções 6.6.2 a 6.6.13. Na subsecção 6.6.14 é efectuado um breve resumo dos resultados obtidos.

6.6.1 Configuração Experimental

Nesta avaliação são comparados os desempenhos dos métodos de combinação de agrupamentos de dados Acumulação de Evidências (EAC - *Evidence Accumulation*), CEAC, CEAC*Boost* e o método de optimização da Médias de Consistência dos Grupos com Penalização de Violação (MCGPV). Tal como na avaliação da secção anterior, foram gerados conjuntos de restrições na forma descrita na secção 6.4. Para todos os métodos de combinação de agrupamentos de dados, foi especificado que o número de grupos do agrupamento de consenso seria o número "real" de grupos dos conjuntos de dados e foram construídos conjuntos de restrições com 10, 20, 50, 100 e 200 restrições. Nas próximas subsecções, são apresentados para cada conjunto de dados os melhores resultados obtidos e os resultados médios resultantes de 20 repetições em cada configuração experimental. Os conjuntos de agrupamentos de dados a combinar foram construídos

usando o algoritmo K-médias, variando aleatoriamente o número de grupos a obter no intervalo $K \in [10, 30]$ com o intuito de criar diversidade, tendo sido gerados 50 agrupamentos de dados para formar cada conjunto de agrupamentos de dados a combinar. No método de combinação de agrupamentos EAC foram usados os algoritmos de agrupamento de dados hierárquicos Ligação Simples (SL - *Single Link*) e Ligação Completa (CL - *Complete Link*) para se extrair da matriz de co-associações os agrupamentos de dados de consenso. Nos métodos CEAC e CEAC*Boost* os algoritmos de agrupamento de dados hierárquicos usados para a extracção do agrupamento de dados de consenso foram o CCL e a versão restringida do algoritmo de Ligação Simples. O número de vizinhos mais semelhantes no método CEAC*Boost* foi definido a 10 vizinhos. No método MCGPV, foi definido como critério de paragem 100 iterações, o tamanho da população a 20, a probabilidade de cruzamento a 80%, a probabilidade de mutação a 1% e a população inicial foi obtida usando o algoritmo de agrupamento de dados K-médias. Note-se que não são apresentados resultados para o método de combinação de agrupamentos de dados que usa o Cop-COBWEB, já que o método é muito lento, devido às suas características recursivas, não tendo sido obtidos resultados suficientes para este método ser incluído nesta avaliação.

6.6.2 Resultados dos Métodos de Combinação de Agrupamentos para o Conjunto de Dados *Bars*

Método	10	20	50	100	200
EAC SL	**99.5%**	**99.5%**	**99.5%**	99.5%	99.5%
EAC CL	53.93%	53.93%	53.93%	53.93%	53.93%
CEAC SL	**99.5%**	**99.5%**	**99.5%**	**100%**	**100%**
CEAC CL	85%	94%	**99.5%**	99.5%	**100%**
CEACBoost SL	98.75%	98.75%	99%	99%	99.5%
CEACBoost CL	98.75%	98.75%	99%	99%	99.75%
MCGPV	99.25%	99.25%	99.25%	99.5%	99.25%

Tabela 6.13: Valores máximos de iC para métodos de combinação de agrupamentos no conjunto de dados *Bars*.

A tabela 6.13 apresenta os melhores resultados obtidos pelos métodos de combinação de agrupamentos de dados, com e sem restrições, para o conjunto de dados *Bars*. Os resultados são bastante satisfatórios, sendo na sua generalidade superiores a 98.75% de iC. Só o método EAC, usando o algoritmo CL para extrair o agrupamento de dados de consenso, obteve um desempenho máximo fraco, com apenas 53.93% de iC. O CEAC, usando tanto o SL como o CL para extrair o agrupamento de dados de consenso, foi o único método que conseguiu agrupar correctamente a totalidade dos objectos deste conjunto de dados. A figura 6.17 apresenta os resultados médios obtidos pelos mesmos métodos de combinação de agrupamentos de dados. Como se pode verificar, o método MCGPV teve um comportamento bastante estável, sendo o método com melhor desempenho médio em todos os conjuntos de restrições, seguido pelos métodos CEAC CL e CEAC*Boost* CL, que obtiveram resultados claramente superiores aos resultados do método EAC, tanto com o uso do SL como com o uso do CL.

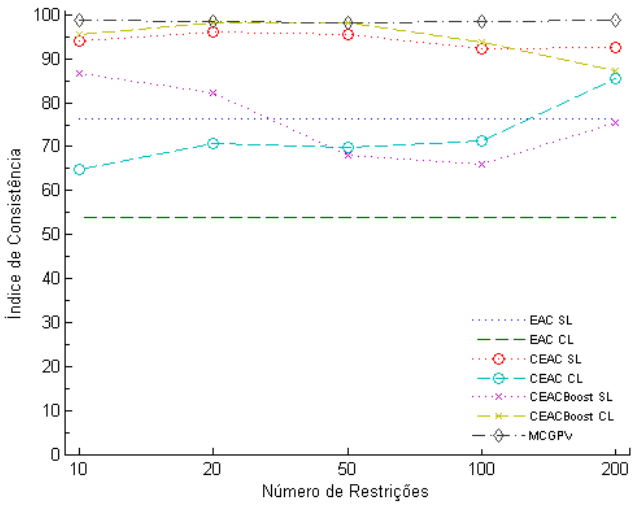

Figura 6.17: Resultados dos métodos de combinação de agrupamentos para o conjunto de dados *Bars* - Valores médios do índice de Consistência.

6.6.3 Resultados dos Métodos de Combinação de Agrupamentos para o Conjunto de Dados *Cigar*

Método	10	20	50	100	200
EAC SL	**100%**	**100%**	**100%**	**100%**	**100%**
EAC CL	43.3%	43.3%	43.3%	43.3%	43.3%
CEAC SL	90%	**100%**	**100%**	**100%**	**100%**
CEAC CL	62.8%	67.2%	83.2%	**100%**	**100%**
CEACBoost SL	89.2%	98.4%	98.4%	**100%**	**100%**
CEACBoost CL	79.6%	86.4%	98.8%	**100%**	**100%**
MCGPV	98.4%	98%	98.4%	98.4%	99.2%

Tabela 6.14: Valores máximos de iC para métodos de combinação de agrupamentos no conjunto de dados *Cigar*.

Os resultados dos métodos de combinação de agrupamentos de dados, com e sem restrições, para o conjunto de dados *Cigar*, são apresentados na tabela 6.14. O método EAC usando o SL para extrair o agrupamento de consenso alcançou 100% de iC. No entanto, o mesmo método obteve apenas 43.3% de iC usando o CL para o mesmo efeito. Relativamente aos métodos de combinação de agrupamento de dados que incorporam restrições, todos os métodos alcançam 100% de iC, com a excepção do MCGPV que obtêm como melhor resultado 99.2% de iC. Nesta tabela é possível verificar que com o aumento do número de restrições, a qualidade do agrupamento de dados de consenso melhora progressivamente. O mesmo se pode concluir analisando a figura 6.18, que apresenta os resultados médios obtidos para o conjunto de dados *Cigar*. O método EAC SL obtém o melhor desempenho, alcançando sempre 100% de iC, independentemente do número de restrições. No entanto, os métodos que usam restrições aproximam-se desse valor quando são usados conjuntos com 200 restrições.

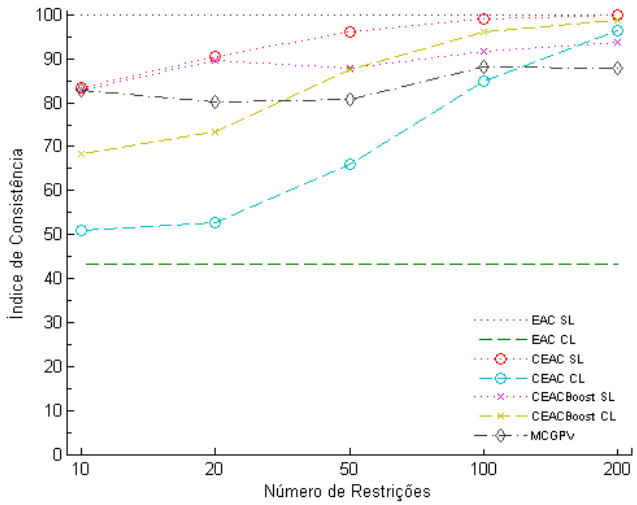

Figura 6.18: Resultados dos métodos de combinação de agrupamentos para o conjunto de dados *Cigar* - Valores médios do índice de Consistência.

6.6.4 Resultados dos Métodos de Combinação de Agrupamentos para o Conjunto de Dados *Spiral*

Método	10	20	50	100	200
EAC SL	97.2%	97.2%	97.2%	97.2%	97.2%
EAC CL	53.05%	53.05%	53.05%	53.05%	53.05%
CEAC SL	**100%**	**100%**	**100%**	**100%**	**100%**
CEAC CL	68.5%	67%	75.5%	88.5%	**100%**
CEACBoost SL	57.5%	57%	62%	73%	92.5%
CEACBoost CL	68%	65%	67.5%	75.5%	73%
MCGPV	64.5%	68%	69%	65%	73.5%

Tabela 6.15: Valores máximos de iC para métodos de combinação de agrupamentos no conjunto de dados *Spiral*.

No conjunto de dados *Spiral*, apenas os métodos EAC e CEAC, usando o SL para extrair o agrupamento de dados de consenso, conseguiram descobrir correctamente os dois grupos em forma de espiral, alcançando como melhores resultados 97.2% e 100% de iC, respectivamente, como mostra a tabela 6.15. Destacam-se também os métodos CEAC CL e CEAC*Boost* SL, que obtiveram respectivamente 100% de iC e 92.5% de iC, em conjuntos com 200 restrições. Os melhores desempenhos dos restantes métodos de combinação de agrupamentos de dados são decepcionantes, variando entre 53.05% de iC no método EAC CL e 88.5% no método CEAC CL com 100 restrições. Os resultados médios para este conjunto de dados encontram-se ilustrados na figura 6.19. Como se pode verificar, com a excepção do método CEAC CL, os resultados dos métodos de combinação de agrupamentos de dados, com e sem restrições, não é satisfatório. No entanto, com o aumento do número de restrições, a qualidade dos agrupamentos de consenso melhora significaticamente.

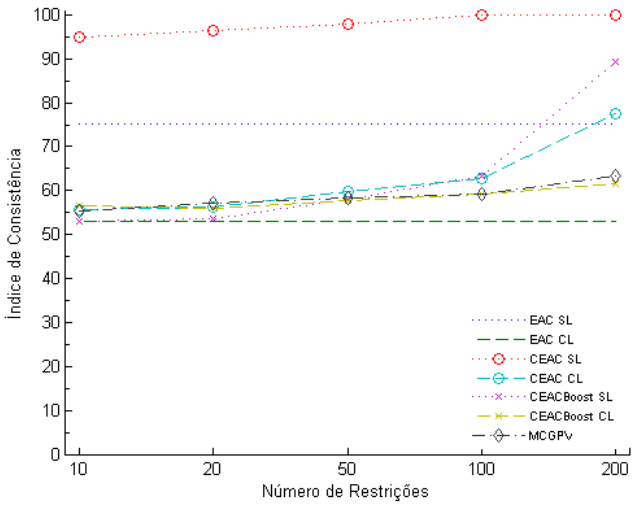

Figura 6.19: Resultados dos métodos de combinação de agrupamentos para o conjunto de dados *Spiral* - Valores médios do índice de Consistência.

6.6.5 Resultados dos Métodos de Combinação de Agrupamentos para o Conjunto de Dados *Half Rings*

Método	10	20	50	100	200
EAC SL	**99.8%**	**99.8%**	99.8%	99.8%	99.8%
EAC CL	45.68%	45.68%	45.68%	45.68%	45.68%
CEAC SL	**99.8%**	**99.8%**	99.8%	**100%**	**100%**
CEAC CL	71.8%	83%	**100%**	**100%**	**100%**
CEACBoost SL	72.8%	77.2%	59.8%	39.8%	47%
CEACBoost CL	80.6%	80.6%	80.4%	78.8%	81%
MCGPV	80%	80.4%	78%	80.8%	83.4%

Tabela 6.16: Valores máximos de iC para métodos de combinação de agrupamentos no conjunto de dados *Half Rings*.

No conjunto de dados *Half Rings*, apenas o método CEAC obtém como melhor resultado 100% de iC, como mostra a tabela 6.16 que apresenta os melhores resultados obtidos para o conjunto de dados. O método EAC obtém como melhor resultado 99.8% de iC e mais nenhum dos restantes métodos de combinação de agrupamentos de dados, com e sem restrições, alcançaram desempenhos semelhantes. O método EAC CL foi o método com pior desempenho, alcançando apenas 45.68% de iC como melhor resultado. Na figura 6.20 são apresentados os resultados médios obtidos pelos métodos de combinação de agrupamentos de dados para o mesmo conjunto de dados. Os métodos EAC e CEAC, usando o algoritmo de agrupamento de dados hierárquico SL para extrair o agrupamento de dados de consenso, obtêm os melhores desempenhos médios, sendo esses resultados bastante superiores aos resultados dos restantes métodos. Apenas o método CEAC CL se aproxima desses resultados e somente quando os conjuntos de restrições são compostos por 200 elementos.

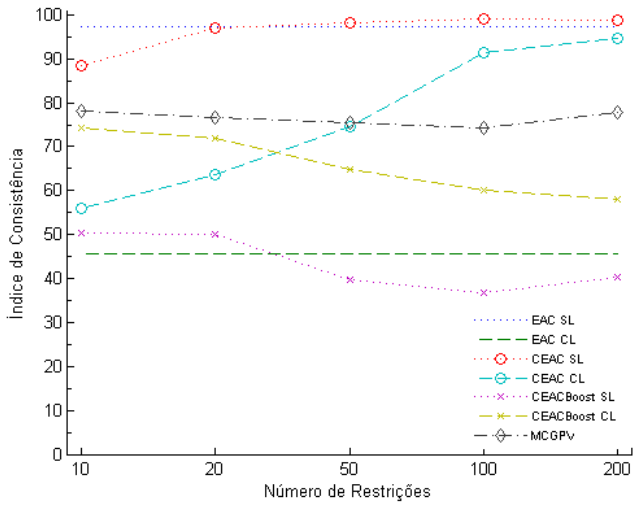

Figura 6.20: Resultados dos métodos de combinação de agrupamentos para o conjunto de dados *Half Rings* - Valores médios do índice de Consistência.

6.6.6 Resultados dos Métodos de Combinação de Agrupamentos para o Conjunto de Dados *Iris*

Método	10	20	50	100	200
EAC SL	73.60%	73.60%	73.60%	73.60%	73.60%
EAC CL	59.72%	59.72%	59.72%	59.72%	59.72%
CEAC SL	**96**%	**96**%	**98**%	**98.67**%	**100**%
CEAC CL	84.67%	94.67%	97.33%	**98.67**%	99.33%
CEACBoost SL	90.67%	90.67%	90%	82%	98.67%
CEACBoost CL	90%	91.33%	92%	95.33%	98%
MCGPV	93.33%	91.33%	96.67%	96.67%	99.33%

Tabela 6.17: Valores máximos de iC para métodos de combinação de agrupamentos no conjunto de dados *Iris*.

No conjunto de dados *Iris*, os métodos de combinação de agrupamentos de dados que incorporam restrições têm desempenhos claramente superiores aos resultados do método EAC, como mostra a tabela 6.17. O método CEAC SL foi o que obteve os melhores desempenhos, alcançando 100% de iC em conjuntos de 200 restrições. Os restantes métodos de combinação de agrupamentos de dados com restrições obtiveram resultados próximos, principalmente nos casos em que o número de restrições é elevado. Observando a figura 6.21, que apresenta os resultados médios obtidos para o conjunto de dados *Iris*, pode-se verificar que os resultados médios do método EAC são o que possuem menor qualidade e que os métodos MCGPV e CEAC SL têm resultados médios semelhantes, sendo estes os dois métodos com melhor desempenho. Realça-se o facto de que apenas o método CEAC CL obteve um desempenho médio inferior ao EAC e apenas com conjuntos de 10 restrições.

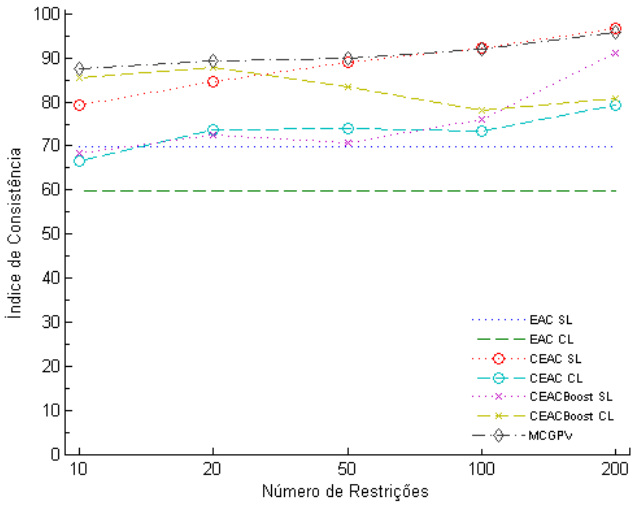

Figura 6.21: Resultados dos métodos de combinação de agrupamentos para o conjunto de dados *Iris* - Valores médios do índice de Consistência.

6.6.7 Resultados dos Métodos de Combinação de Agrupamentos para o Conjunto de Dados *Breast Cancer*

Método	10	20	50	100	200
EAC SL	95.05%	95.05%	95.05%	95.05%	95.05%
EAC CL	62.75%	62.75%	62.75%	62.75%	62.75%
CEAC SL	**97.36%**	**97.07%**	**97.51%**	**97.36%**	**97.95%**
CEAC CL	92.97%	**97.07%**	97.07%	96.34%	97.51%
CEACBoost SL	96.34%	96.34%	96.49%	67.79%	68.67%
CEACBoost CL	96.49%	96.34%	96.49%	96.63%	67.2%
MCGPV	92.24%	92.24%	92.09%	93.7%	93.56%

Tabela 6.18: Valores máximos de iC para métodos de combinação de agrupamentos no conjunto de dados *Breast Cancer*.

A tabela 6.18 apresenta os melhores resultados obtidos pelos métodos de combinação de agrupamentos de dados, com e sem restrições, para o conjunto de dados *Breast Cancer*. Com a excepção do método EAC CL que alcançou apenas 62.75%, todos os métodos de combinação obtiveram valores de iC acima dos 92%. Neste conjunto de dados, destaca-se uma vez mais o desempenho do método CEAC, usando o SL para efectuar a extracção do agrupamento de dados de consenso, obtendo em todos os conjuntos de restrições os melhores valores absolutos de iC. Na figura 6.22 são apresentados os resultados médios obtidos para o conjunto de dados *Breast Cancer*. Os métodos CEAC SL e MCGPV são os únicos métodos com resultados médios sempre superiores a 85% de iC. O método EAC CL tem novamente o pior desempenho com cerca de 63% de iC e o método EAC SL obtém valores médios de 84% de iC, o que permite concluir que a escolha do algoritmo para extrair o agrupamento de dados de consenso é especialmente importante nos métodos não supervisionados, já que nos métodos que usam restrições não existem diferenças tão significativas.

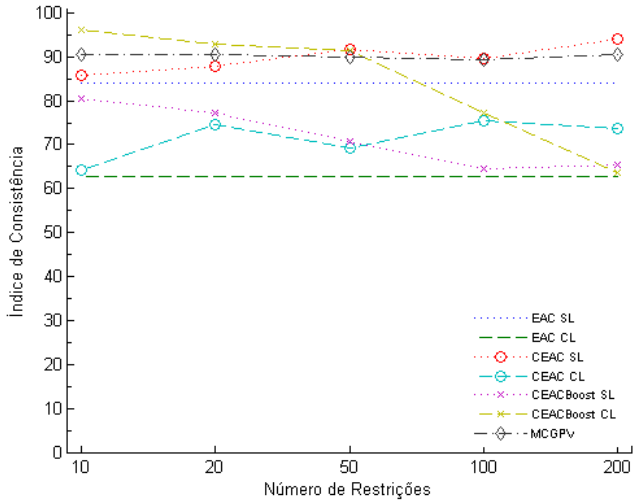

Figura 6.22: Resultados dos métodos de combinação de agrupamentos para o conjunto de dados *Breast Cancer* - Valores médios do índice de Consistência.

6.6.8 Resultados dos Métodos de Combinação de Agrupamentos para o Conjunto de Dados *Log Yeast*

Método	10	20	50	100	200
EAC SL	43.7%	43.7%	43.7%	43.7%	43.7%
EAC CL	38.54%	38.54%	38.54%	38.54%	38.54%
CEAC SL	**45.31%**	**52.6%**	**56.51%**	**56.77%**	**55.21%**
CEAC CL	42.19%	49.22%	45.05%	53.13%	47.4%
CEACBoost SL	38.8%	40.1%	41.41%	45.31%	45.57%
CEACBoost CL	39.84%	39.06%	35.16%	37.76%	40.62%
MCGPV	33.33%	32.29%	31.25%	32.29%	34.9%

Tabela 6.19: Valores máximos de iC para métodos de combinação de agrupamentos no conjunto de dados *Log Yeast*.

Os melhores resultados obtidos pelos métodos de combinação de agrupamentos de dados, com e sem restrições, para o conjunto de dados *Log Yeast* são apresentados na tabela tabela 6.19. Os desempenhos de todos os métodos de combinação de agrupamentos de dados são de fraca qualidade, tendo apenas o método CEAC CL obtido valores de iC superiores a 50%. Neste conjunto de dados, o uso de restrições não implicou o aumento da qualidade dos agrupamentos de dados de consenso, tendo o método MCGPV alcançando os piores valores de iC, tanto nos valores máximos (tabela 6.19) como nos valores médios (figura 6.23).

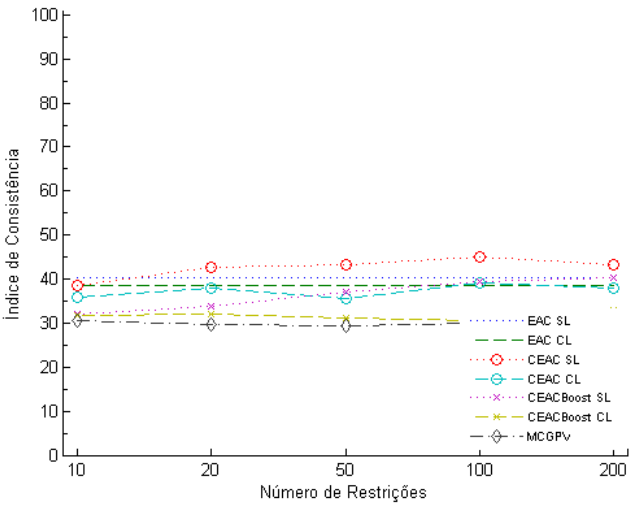

Figura 6.23: Resultados dos métodos de combinação de agrupamentos para o conjunto de dados LogYeast - Valores médios do índice de Consistência.

6.6.9 Resultados dos Métodos de Combinação de Agrupamentos para o Conjunto de Dados *Std Yeast*

Método	10	20	50	100	200
EAC SL	56.98%	56.98%	56.98%	56.98%	56.98%
EAC CL	46.59%	46.59%	46.59%	46.59%	46.59%
CEAC SL	60.94%	61.72%	63.02%	64.58%	70.05%
CEAC CL	49.22%	54.95%	49.48%	57.55%	51.04%
CEACBoost SL	59.64%	66.41%	65.89%	61.2%	52.86%
CEACBoost CL	72.66%	68.49%	65.89%	61.72%	60.68%
MCGPV	**73.7%**	**73.18%**	**71.09%**	**69.79%**	**71.61%**

Tabela 6.20: Valores máximos de *iC* para métodos de combinação de agrupamentos no conjunto de dados *Std Yeast*.

A tabela 6.20 apresenta os melhores resultados obtidos pelos métodos de combinação de agrupamentos de dados, com e sem restrições, para o conjunto de dados *Std Yeast*. Nesta versão do conjunto de dados *Yeast Cell Cycle* os resultados são significativamente superiores, comparativamente aos resultados do conjunto de dados *Log Yeast*. Neste conjunto de dados, apenas o método CEAC CL obtêm melhores resultados inferiores aos melhores resultados do método EAC. Os métodos EAC alcançam apenas 56.98% e 46.59% de *iC* usando, respectivamente, os algoritmos SL e CL para extrair o agrupamento de dados de consenso. Destaca-se o desempenho do método MCGPV que obteve nos conjuntos de 10, 20, 50, 100 e 200 restrições os melhores resultados, sendo os valores de *iC* 73.7%, 73.18%, 71.09%, 69.79% e 71.61%, respectivamente. Na figura 6.24 são apresentados os resultados médios obtidos para este conjunto de dados. Uma vez mais, o método MCGPV obtém os melhores desempenhos médios, sendo o CEAC CL o método com pior desempenho.

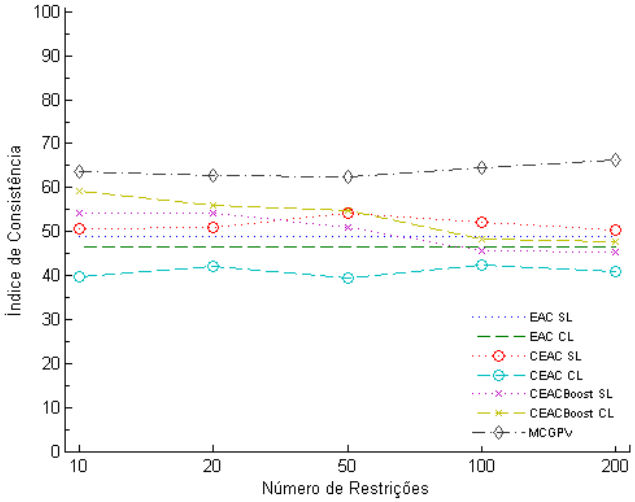

Figura 6.24: Resultados dos métodos de combinação de agrupamentos para o conjunto de dados *Std Yeast* - Valores médios do índice de Consistência.

6.6.10 Resultados dos Métodos de Combinação de Agrupamentos para o Conjunto de Dados *Optdigits*

Método	10	20	50	100	200
EAC SL	65.14%	65.14%	65.14%	65.14%	65.14%
EAC CL	56.81%	56.81%	56.81%	56.81%	56.81%
CEAC SL	39.1%	49.2%	59.1%	75.4%	**90.3**%
CEAC CL	73.6%	73.5%	70.6%	77%	78.5%
CEACBoost SL	39.2%	49.2%	58.4%	75.3%	75.7%
CEACBoost CL	71.7%	75.7%	73.7%	**84**%	77%
MCGPV	**83.8**%	**83.2**%	**82.7**%	82.2%	83.9%

Tabela 6.21: Valores máximos de iC para métodos de combinação de agrupamentos no conjunto de dados *Optdigits*.

No conjunto de dados *Optdigits*, o método de combinação de agrupamentos de dados que obteve o melhor valor de iC foi o CEAC SL com 90.3%, usando 200 restrições, como mostra a tabela 6.21. No entanto, o método MCGPV obteve em todos os conjuntos de restrições valores na ordem dos 83%. Os melhores resultados do método EAC foram 65.14% de iC e 56.81% de iC para os algoritmos SL e CL, respectivamente. A figura 6.25 ilustra os resultados médios obtidos pelos métodos de combinação de agrupamentos de dados, com e sem restrições, para o conjunto de dados *Optdigits*. Como mostra a figura 6.25 o método MCGPV é bastante estável, obtendo sempre valores médios muito próximos de 80% de iC. O método EAC SL tem como resultado médio um valor próximo dos 56% de iC e o EAC CL um pouco inferior a 55%. Neste conjunto de dados pode-se verificar que os métodos CEAC SL e CEAC*Boost* SL melhoraram muito significativamente a qualidade dos agrupamentos de dados de consenso com o aumento do número de restrições.

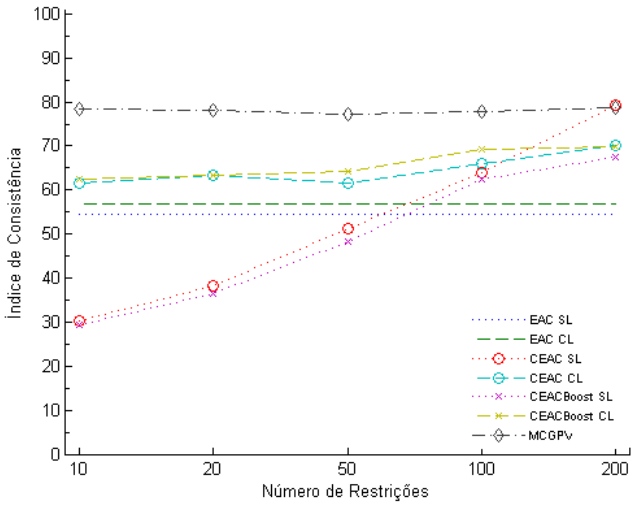

Figura 6.25: Resultados dos métodos de combinação de agrupamentos para o conjunto de dados *Optdigits* - Valores médios do índice de Consistência.

6.6.11 Resultados dos Métodos de Combinação de Agrupamentos para o Conjunto de Dados *Glass*

Método	10	20	50	100	200
EAC SL	51.4%	51.4%	51.4%	51.4%	51.4%
EAC CL	39.42%	39.42%	39.42%	39.42%	39.42%
CEAC SL	**59.81%**	**62.15%**	**65.89%**	**64.02%**	**76.17%**
CEAC CL	42.99%	53.74%	55.61%	55.14%	56.07%
CEACBoost SL	53.27%	52.34%	51.87%	57.94%	70.09%
CEACBoost CL	53.74%	54.67%	52.34%	62.15%	74.3%
MCGPV	52.8%	48.6%	51.4%	45.79%	48.13%

Tabela 6.22: Valores máximos de iC para métodos de combinação de agrupamentos no conjunto de dados *Glass*.

A tabela 6.22 apresenta os melhores resultados obtidos pelos métodos de combinação de agrupamentos de dados, com e sem restrições, no conjunto de dados *Glass*. Nenhum dos métodos obteve bons resultados, tendo no entanto, o método de combinação de agrupamentos de dados CEAC SL obtido o valor máximo de 76.17% usando conjuntos de 50 restrições. O método com pior desempenho foi o EAC CL, alcançando apenas 39.42% de iC. Na figura 6.26 estão ilustrados os resultados médios para o conjunto de dados *Glass*. O método EAC CL tem o pior desempenho médio, com cerca de 40% de iC. Os resultados dos restantes métodos de combinação de agrupamentos de dados variam geralmente entre esse valor e 50%, com a excepção do método CEAC*Boost* nos conjuntos com 100 e 200 restrições.

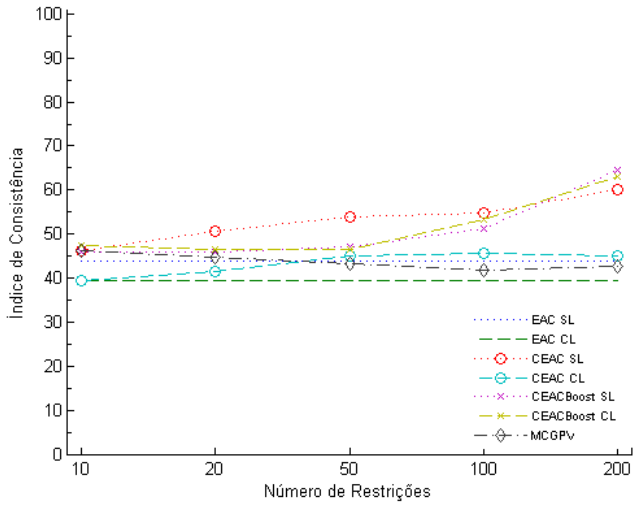

Figura 6.26: Resultados dos métodos de combinação de agrupamentos para o conjunto de dados *Glass* - Valores médios do índice de Consistência.

6.6.12 Resultados dos Métodos de Combinação de Agrupamentos para o Conjunto de Dados *Wine*

Método	10	20	50	100	200
EAC SL	**72.47%**	**72.47%**	72.47%	72.47%	72.47%
EAC CL	51.03%	51.03%	51.03%	51.03%	51.03%
CEAC SL	**72.47%**	70.79%	65.73%	64.04%	73.6%
CEAC CL	61.8%	62.36%	59.55%	65.73%	69.66%
CEACBoost SL	70.79%	67.98%	67.42%	69.1%	**86.52%**
CEACBoost CL	**72.47%**	**72.47%**	71.35%	69.1%	76.4%
MCGPV	71.35%	71.91%	**73.03%**	**73.03%**	76.97%

Tabela 6.23: Valores máximos de iC para métodos de combinação de agrupamentos no conjunto de dados *Wine*.

Os melhores resultados obtidos pelos métodos de combinação de agrupamentos de dados, com e sem restrições, para o conjunto de dados *Wine* são apresentados na tabela 6.23. O CEAC*Boost*, usando o algoritmo de agrupamento de dados hierárquico SL para extrair o agrupamento de dados de consenso, foi o método que obteve o valor máximo de iC com 86.52%. Os restantes métodos de combinação de agrupamentos alcançaram desempenhos máximos inferiores em pelo menos 9%, tendo o método EAC CL obtido apenas 51.03% de iC. Relativamente aos resultados médios apresentados na figura 6.27, o método EAC SL obtém quase sempre os melhores resultados médios, com cerca de 70% de iC, valor apenas superado pelo método MCGPV usando conjuntos de 200 restrições. Os métodos CEAC*Boost* e CEAC SL evidenciam uma tendência contrária à esperada, decrescendo a qualidade dos agrupamentos de dados de consenso com o aumento do número de restrições. Estas tendências foram apenas contrariadas usando conjuntos com 200 restrições.

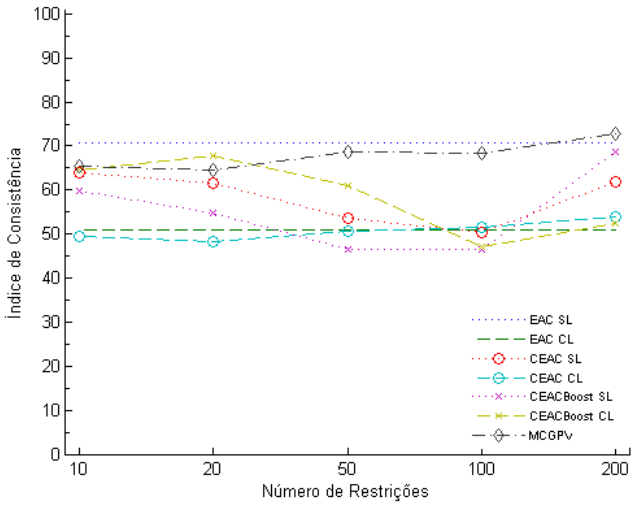

Figura 6.27: Resultados dos métodos de combinação de agrupamentos para o conjunto de dados *Wine* - Valores médios do índice de Consistência.

6.6.13 Resultados dos Métodos de Combinação de Agrupamentos para o Conjunto de Dados *Image Segmentation*

Método	10	20	50	100	200
EAC SL	29.12%	29.12%	29.12%	29.12%	29.12%
EAC CL	42.41%	42.41%	42.41%	42.41%	42.41%
CEAC SL	42.86%	51.65%	55.71%	**65.28**%	**67.49**%
CEAC CL	52.51%	40.52%	46.02%	54.29%	52.68%
CEACBoost SL	42.86%	42.86%	43.94%	42.51%	36.67%
CEACBoost CL	**58.79**%	55.67%	**55.84**%	54.59%	65.84%
MCGPV	56.28%	**58.66**%	54.42%	54.42%	52.9%

Tabela 6.24: Valores máximos de *iC* para métodos de combinação de agrupamentos no conjunto de dados *Image Segmentation*.

A tabela 6.24 apresenta os melhores resultados obtidos para o último conjunto de dados usado nesta avaliação, o conjunto de dados *Image Segmentation*. Os resultados obtidos para todos os métodos de combinação de agrupamentos de dados são de pouca qualidade, tendo o método CEAC SL obtido o melhor valor de *iC* com 65.28% e 67.49% para conjuntos com 100 e 200 restrições. O método com pior resultado máximo foi o EAC com 29.12% de *iC* e 42.41% de *iC* com os algoritmos SL e CL, respectivamente. Relativamente aos resultados médios, ilustrados na figura 6.28, verifica-se novamente que o método EAC SL obteve um mau desempenho médio, com valor de *iC* próximo de 28%. O método EAC CL obteve para este conjunto de dados o resultado médio de cerca de 42% de *iC*, sendo no entanto esse valor geralmente inferior aos valores de *iC* obtidos pelos métodos de combinação de agrupamentos de dados que usam restrições, com a excepção do CEAC*Boost* SL. O método CEAC SL é uma vez mais o método com melhor desempenho médio, evidenciado claramente uma tendência crescente na qualidade do agrupamento de dados de consenso com o aumento do número de restrições.

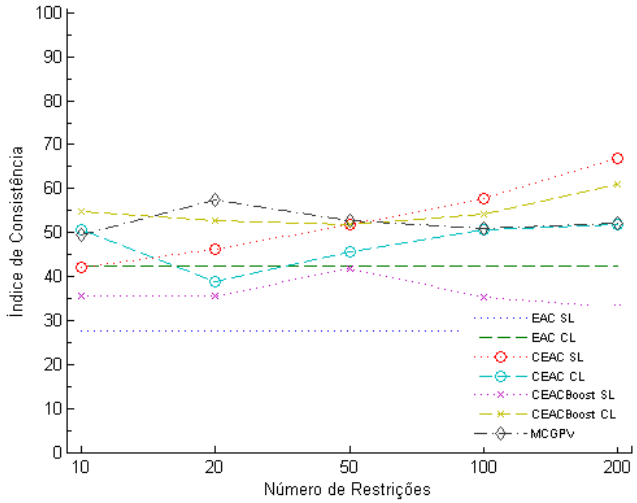

Figura 6.28: Resultados dos métodos de combinação de agrupamentos para o conjunto de dados *Image Segmentation* - Valores médios do índice de Consistência.

6.6.14 Resumo dos Resultados dos Métodos de Combinação de Agrupamentos de Dados

Com a análise dos resultados dos métodos de combinação de agrupamentos de dados, concluiu-se que os métodos que incorporam restrições obtêm geralmente melhores resultados que os métodos de combinação de agrupamentos de dados não supervisionados. De todos os métodos de combinação de agrupamentos de dados apresentados, com e sem restrições, o método CEAC usando o algoritmo de agrupamento de dados hierárquico SL obteve mais vezes o valor máximo do índice de Consistência, mais precisamente, em 10 dos 12 conjuntos de dados. O método EAC igualou o valor máximo do índice de Consistência apenas num conjunto de dados (*Cigar*), nunca tendo obtido resultados máximos superiores a todos os métodos de combinação de agrupamento de dados que usam restrições. Nesta avaliação, destacam-se também os métodos CEAC CL, CEAC*Boost* SL e MCGPV, tendo o primeiro método obtido o valor máximo de iC em três dos conjuntos de dados e os restantes métodos em dois dos conjuntos de dados.

6.7 Sumário

Neste capítulo foram apresentados dois estudos comparativos, com o intuito de comparar o desempenho de alguns dos algoritmos de agrupamentos de dados com restrições, descritos no capítulo 4 desta dissertação, e de avaliar os métodos de combinação de agrupamentos de dados com restrições propostos no capítulo 5, tendo sido utilizados 12 conjuntos de dados distintos para o efeito. Com estas avaliações conclui-se que os algoritmos de agrupamento de dados com restrições e que os métodos de combinação de agrupamentos de dados com restrições obtêm geralmente melhores desempenhos que os algoritmos de agrupamento de dados e métodos de combinação de agrupamentos de dados não supervisionados. Observou-se também que em alguns (poucos) conjuntos de dados, o uso de restrições originou diminuição de qualidade no agrupamento de dados. Comparando os resultados dos algoritmos de agrupamento de dados, com e sem restrições, com os métodos de combinação de agrupamentos de dados, com e sem restrições, em 6 conjuntos dos dados os algoritmos de agrupamento de dados com restrições obtêm valores de iC máximos superiores aos valores máximos obtidos pelos métodos de combinação. No entanto, nos restantes 6 conjuntos de dados acontece o inverso, já que os métodos de combinação de agrupamentos de dados com restrições obtêm valores máximos de iC superiores aos valores máximos obtidos pelos algoritmos de agrupamentos de dados.

Capítulo 7

Conclusões

7.1 Resumo

Nesta dissertação investigou-se o tema do agrupamento de dados com restrições, um campo da aprendizagem semi-supervisionada actualmente em grande actividade, em que é utilizada informação *a priori* na forma de restrições. Esta informação pode ser obtida usando supervisão humana ou fontes automáticas e é incorporada no agrupamento de dados com o intuito de proporcionar soluções adaptadas a tarefas ou interesses específicos.

Existem vários níveis de restrições, desde as restrições globais que se aplicam a todo o conjunto de dados, passando pelas restrições ao nível dos grupos e ao nível dos atributos, até restrições a níveis mais específicos, as restrições ao nível dos objectos. Esta dissertação focalizou-se nas restrições ao nível dos objectos, mais particularmente, nas restrições do tipo rotulação parcial e nas relações entre pares de objectos, dado que grande parte das restrições dos níveis superiores podem ser expressas na forma destes dois últimos tipos de restrições.

Foram apresentados os principais algoritmos de agrupamento de dados que usam restrições, organizando-os em cinco categorias distintas: *Restrições Invioláveis*, em que os algoritmos de agrupamento de dados com restrições garantem que as soluções encontradas satisfazem completamente todas as restrições; *Restrições na Forma de Rótulos*, em que a um subconjunto dos objectos de dados se encontram associados rótulos, servindo normalmente esses rótulos para inicializar os centros dos grupos dos algoritmos de agrupamento de dados de partição. Este tipo de restrições pode ser também utilizado para influenciar a fase de atribuição dos objectos de dados aos grupos cujos centros sejam mais próximos; *Penalização de Violações de Restrições*, em que para além de considerar as distâncias entre os objectos de dados e os respectivos centros de grupo, é geralmente adaptada uma função-objectivo para penalizar a violação de restrições, não sendo necessário que todas as restrições sejam satisfeitas; *Edição de Distância*, em que se aprende uma medida de distância com o intuito de generalizar as restrições entre objectos de dados ao nível do espaço dos atributos de dados, propagando as restrições entre pares de objectos a outros objectos próximos, que podem não ter sido incluídos nos conjuntos de restrições; e,

finalmente, *Modificação do Processo de Geração*, em que se pressupõe que os objectos de dados foram gerados segundo um modelo probabilístico, sendo esse modelo modificado com o objectivo de se considerar relações entre objectos na estimação dos parâmetros do modelo.

Nesta dissertação, foi também abordado o tema da combinação de soluções de algoritmos de aprendizagem, isto é, a combinação de classificadores de dados e a combinação de agrupamentos de dados, com o objectivo de se estudar quais as principais vantagens e desvantagens da combinação de soluções em ambas as aprendizagens supervisionada e não supervisionada. Este estudo foi realizado para, posteriormente, se aplicar a combinação de soluções à aprendizagem semi-supervisionada, mais precisamente, a combinação de agrupamentos de dados usando restrições. Foram propostos quatro métodos distintos para se combinar agrupamentos de dados usando restrições: dois deles são versões modificadas do método de Acumulação de Evidências, em que são usados algoritmos de agrupamento de dados hierárquicos com restrições para extrair o agrupamento de consenso da matriz de co-associações; um método transforma o conjunto de agrupamentos de dados numa nova representação do conjunto de dados, usando em seguida o algoritmo de agrupamento de dados categóricos COP-COBWEB como função de consenso; e um método é baseado na optimização da Medida de Consistência de Grupos com penalização de violações de restrições usando um algoritmo genético para maximizar a função-objectivo.

Finalmente, foram realizados dois estudos tendo em vista a avaliação de vários dos algoritmos de agrupamento e métodos de combinação de agrupamentos de dados com restrições apresentados nesta dissertação, tendo os respectivos desempenhos sido comparados em conjuntos de dados sintéticos e reais. Com o primeiro estudo concluiu-se que os algoritmos de agrupamento de dados que usam restrições obtêm geralmente melhores resultados que os algoritmos de agrupamento de dados não supervisionados. De todos os algoritmos de agrupamento com restrições avaliados, o K-médias Restringido foi o algoritmo com melhor desempenho, tendo obtido em 10 dos 12 conjuntos de dados os valores máximos do índice de Consistência. No segundo estudo, concluiu-se também que o uso de restrições favorece a qualidade do agrupamento de dados de consenso. Neste estudo, o método de combinação de agrupamentos de dados com melhor desempenho foi o CEAC SL, tendo obtido os valores máximos do índice de Consistência em 10 dos 12 conjuntos de dados.

7.2 Objectivos Alcançados

O primeiro objectivo desta dissertação consistiu na revisão do estado da arte do agrupamento de dados com restrições, tendo sido estudados vários tipos de restrições usadas no agrupamento de dados e os principais algoritmos de agrupamento de dados com restrições. Foram também apresentados os conceitos fundamentais das aprendizagens supervisionada, não supervisionada e semi-supervisionada com o intuito de enquadrar o agrupamento de dados com restrições na aprendizagem automática. Foi estudado o tema da combinação de soluções de algoritmos de

aprendizagem, pois foi objectivo desta dissertação o desenvolvimento de métodos de combinação de agrupamentos de dados com restrições.

Foram propostas quatro formas distintas de combinar agrupamentos de dados usando restrições. A primeira proposta (CEAC) baseia-se no método de Acumulação de Evidências [35] diferenciando-se simplesmente na utilização de um algoritmo de agrupamento de dados hierárquico com restrições na extracção do agrupamento de dados de consenso, permitindo desta forma impor restrições de relações entre pares de objectos. A segunda proposta (CEAC*Boost*) expande o método CEAC na construção dos agrupamentos de dados a combinar, sendo os agrupamentos de dados gerados sequencialmente e dando-se maior importância aos objectos mais difíceis de agrupar, tal como nos algoritmos de *Boosting* da aprendizagem supervisionada. Neste método, os agrupamentos de dados a combinar são obtidos usando amostragem com reposição dos dados, tendo cada objecto uma probabilidade diferente de ser incluído nas amostras de dados que produzem os agrupamentos de dados a combinar. Essa probabilidade é inversamente proporcional à confiança de atribuição do objecto aos grupos, que é calculada com base nas similaridades desse objecto aos K vizinhos mais similares, segundo a matriz de co-associações entre objectos. O terceiro método consiste em representar os agrupamentos de dados a combinar num novo conjunto de dados categóricos, em que cada agrupamento de dados a combinar corresponde a um novo meta-atributo de dados. Os valores de cada atributo na nova representação correspondem aos rótulos dos objectos no agrupamento de dados correspondente. Após a nova representação do conjunto de agrupamentos de dados ter sido construída, é aplicado o algoritmo de agrupamento de dados com restrições COP-COBWEB [85] para se obter o agrupamento de dados de consenso, satisfazendo as restrições impostas. Finalmente, o último método proposto consiste na maximização de uma função-objectivo baseada na Medida de Consistência de Grupos [27] (MCG) com penalização de violações de restrições. A MCG mede a similaridade entre os agrupamentos de dados a combinar e o agrupamento de dados alvo que se pretende avaliar. Neste método, é procurado o agrupamento de consenso P^* que maximiza a MCG entre P^* e o conjunto de agrupamentos de dados a combinar e que, simultaneamente, maximize a satisfação das restrições. Para o efeito, a função-objectivo é maximizada usando um algoritmo genético.

Finalmente, o último objectivo desta dissertação consistiu na comparação dos desempenhos de vários algoritmos de agrupamentos de dados que usam restrições, apresentados no capitulo 4, e dos métodos de combinação de agrupamentos de dados com restrições, propostos no capitulo 5, de forma a avaliar a qualidade dos agrupamentos de dados produzidos por cada algoritmo/método, em doze conjuntos de dados diferentes.

7.3 Limitações e Trabalho Futuro

Nesta dissertação, apresentaram-se os principais algoritmos de agrupamento de dados que incorporam restrições. No entanto, existem outros algoritmos de agrupamento de dados com restrições que não puderam ser incluídos neste documento, devido a limitações de espaço. Nem

todos os algoritmos de agrupamento de dados com restrições apresentados foram implementados, o que limitou o estudo comparativo relativo aos algoritmos de agrupamento de dados com restrições. Mesmo considerando apenas os algoritmos de agrupamento de dados com restrições implementados, houve conjuntos de dados em que não foram apresentados resultados, devido às elevadas cardinalidades dos conjuntos de dados e aos elevados custos computacionais dos algoritmos de agrupamento de dados. Relativamente aos métodos de combinação de agrupamentos de dados com restrições, não foram examinados exaustivamente qual a melhor forma de construir os conjuntos de agrupamentos de dados. Na construção dos agrupamentos de dados poder-se-á variar, nomeadamente, o número de grupos dos agrupamentos de dados a combinar, o número de agrupamentos de dados a combinar, o uso ou não de amostragem de dados (com e sem reposição de dados), e o uso de vários algoritmos de agrupamento de dados (usando ou não restrições). Na construção dos conjuntos de restrições foram construídos conjuntos com 10, 20, 50, 100 e 200 relações entre objectos de dados, independentemente do conjunto de dados. Para que se pudesse testar com maior exactidão a influência do número de restrições no desempenho dos algoritmos de agrupamento de dados com restrições e dos métodos de combinação de agrupamentos de dados com restrições, o número de restrições em cada conjunto deveria ser proporcional à cardinalidade de cada conjunto de dados.

Na avaliação dos algoritmos de agrupamento de dados e dos métodos de combinação de agrupamentos de dados que usam restrições, não foi testada a presença de ruído no conjunto de restrições, o que não possibilitou o estudo do comportamento dos algoritmos de agrupamento de dados e dos métodos de combinação de agrupamentos de dados quando uma fracção da informação *a priori* não se encontra correcta. Nas avaliações, apenas um algoritmo de agrupamento de dados e um método de combinação de agrupamentos de dados não supervisionado foram usados como referência para comparação com os algoritmos de agrupamento de dados e dos métodos de combinação de agrupamentos de dados que usam restrições. Seria pertinente comparar os resultados desses algoritmos e métodos com uma maior variedade de algoritmos de agrupamento de dados e de métodos de combinação de agrupamentos de dados não supervisionados para que as conclusões tiradas tivessem um maior fundamento.

Seria interessante usar as restrições de ligação obrigatória e de ligação proibida na construção da matriz de co-associações, tanto no método de combinação de agrupamentos de dados CEAC como no CEAC*Boost*, de forma a influenciar a nova medida de similaridade entre objectos do conjunto de dados, em vez de apenas se utilizar as restrições na extracção do agrupamento de dados de consenso.

No método de combinação de agrupamentos de dados CEAC*Boost*, é dada maior importância aos objectos mais difíceis de agrupar. Para isso é usada amostragem de dados com reposição, em que cada objecto tem uma probabilidade de ser seleccionado inversamente proporcional à confiança da sua atribuição aos grupos no conjunto de agrupamentos de dados. Em vez de se utilizar amostragem de dados com reposição, poder-se-ia utilizar directamente as probabilidades de

cada objecto ser seleccionado, como ponderação de cada objecto nos algoritmos de agrupamento de dados para gerar o conjunto de agrupamento de dados a combinar.

O algoritmo de agrupamento de dados com restrições COP-COBWEB foi proposto para servir de função de consenso para a combinação de agrupamentos de dados. Neste caso, o conjunto de agrupamentos de dados é mapeado numa nova representação do conjunto de dados cujo atributos são categóricos. No entanto, o algoritmo COP-COBWEB é muito lento pelo que devem ser exploradas outras alternativas. Uma possibilidade consiste em representar o conjunto de agrupamentos de dados num conjunto de dados contínuos e em seguida aplicar um dos vários algoritmos de agrupamento com restrições para dados contínuos.

O método de combinação de agrupamentos de dados baseado na optimização da Média da Consistência dos Grupos com penalização de violações demora bastante tempo a encontrar agrupamentos de dados de qualidade aceitável, pelo que, seria interessante usar as restrições para influenciar a pesquisa da melhor solução no espaço de soluções. Para isso, ter-se-ia de encontrar formas para realizar o cruzamento e mutação dos indivíduos da população corrente de forma "inteligente", isto é, favorecer o aparecimento de novos indivíduos que satisfaçam as restrições impostas.

Para além das sugestões atrás apresentadas, na tentativa de superar as limitações encontradas, pretende-se no futuro conduzir o trabalho nas seguintes direcções de investigação:

- Aprendizagem de medidas de distância entre objectos tendo em conta não só os atributos do conjunto de dados mas também as restrições entre objectos;
- Concepção de métodos inteligentes que orientem o utilizador na transformação de conhecimento de domínio em restrições para posterior utilização em algoritmos de agrupamento de dados e métodos de combinação de agrupamentos de dados;
- Desenvolvimento de métodos visuais e interactivos de agrupamento e de ferramentas que apoiem o utilizador na interacção durante o processo de agrupamento.
- Aplicação dos algoritmos de agrupamento de dados com restrições e dos métodos de combinação de agrupamentos de dados com restrições a vários problemas reais, em particular, na estruturação de conteúdos multimédia existentes na *Web* e de correspondência electrónica.

Bibliografia

[1] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, "Automatic subspace clustering of high dimensional data for data mining applications," *SIGMOD Rec.*, vol. 27, no. 2, pp. 94–105, 1998. 12

[2] M. Al-Razgan and C. Domeniconi, "Weighted clustering ensembles," in *SDM*, 2006. 72

[3] E. Alpaydin and C. Kaynak, UCI Machine Learning Repository [http://www.ics.uci.edu/ mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science., 1998. 87

[4] E. Anderson, "The irises of the gaspe peninsula," *Bulletin of the American Iris Society*, vol. 59, pp. 2–5, 1935. 87

[5] S. Basu, A. Banerjee, and R. Mooney, "Semi-supervised clustering by seeding," 2002. [Online]. Available: citeseer.ist.psu.edu/basu02semisupervised.html 23, 31, 52

[6] S. Basu, M. Bilenko, and R. Mooney, "Comparing and unifying search-based and similarity-based approaches to semi-supervised clustering," 2003. [Online]. Available: citeseer.ist.psu.edu/article/basu03comparing.html 42

[7] S. Basu, A. Banerjee, and R. J. Mooney, "Active semi-supervision for pairwise constrained clustering," in *In Proceedings of the 2004 SIAM International Conference on Data Mining (SDM-04*, 2004, pp. 333–344. 25, 32, 34

[8] S. Basu, M. Bilenko, and R. J. Mooney, "A probabilistic framework for semi-supervised clustering," in *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining.* New York, NY, USA: ACM, 2004, pp. 59–68. 49, 54

[9] E. Bauer and R. Kohavi, "An empirical comparison of voting classification algorithms: Bagging, boosting, and variants," *Mach. Learn.*, vol. 36, no. 1-2, pp. 105–139, 1999. 61

[10] K. P. Bennett and A. Demiriz, "Semi-supervised support vector machines," in *Proceedings of the 1998 conference on Advances in neural information processing systems II.* Cambridge, MA, USA: MIT Press, 1999, pp. 368–374. 16

[11] M. Bilenko, S. Basu, and R. J. Mooney, "Integrating constraints and metric learning in semi-supervised clustering," in *ICML '04: Proceedings of the twenty-first international conference on Machine learning.* New York, NY, USA: ACM, 2004, p. 11. 46

[12] C. Blake, UCI Machine Learning Repository [http://www.ics.uci.edu/ mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science., 1998. 88

[13] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in

COLT' 98: Proceedings of the eleventh annual conference on Computational learning theory. New York, NY, USA: ACM, 1998, pp. 92–100. 14

[14] U. Bodenhofer, "Genetic algorithms: Theory and applications," Lecture Note Third Edition — Winter 2003/2004, 2003. 79

[15] P. S. Bradley, K. P. Bennett, and A. Demiriz, "Constrained k-means clustering," 2000. [Online]. Available: citeseer.ist.psu.edu/bradley00constrained.html 21

[16] L. Breiman, J. Friedman, R. Olshen, and C. Stone, "Classification and regression trees," Wadsworth, Belmont, CA, 1984. 6, 23

[17] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996. 61

[18] C. Brodley, UCI Machine Learning Repository [http://www.ics.uci.edu/ mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science., 1990. 88

[19] D. G. C. Domeniconi, D. Papadopoulos and S. Ma, "Subspace clustering of high dimensional data," in *SIAM International Conference on Data Mining (SDM)*, 2004. 72

[20] D. Cohn, R. Caruana, and A. McCallum, "Semi-supervised clustering with user feedback," 2003. [Online]. Available: citeseer.ist.psu.edu/cohn03semisupervised.html 24, 25

[21] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines.* Cambridge: Cambridge University Press, 2000. 6

[22] I. Davidson and S. Ravi, "Clustering with constraints feasibility issues and the k-means algorithm," in *2005 SIAM International Conference on Data Mining (SDM'05)*, Newport Beach,CA, 2005, pp. 138–149. 32, 36

[23] A. Demiriz, K. P. Bennett, and M. J. Embrechts, "Semi-supervised clustering using genetic algorithms," in *In Artificial Neural Networks in Engineering (ANNIE-99.* ASME Press, 1999, pp. 809–814. 23

[24] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977. 12, 48, 73

[25] T. G. Dietterich, "Ensemble methods in machine learning," in *MCS '00: Proceedings of the First International Workshop on Multiple Classifier Systems.* London, UK: Springer-Verlag, 2000, pp. 1–15. 57, 58

[26] E. Dimitriadou, A. Weingessel, and K. Hornik, "Voting-merging: An ensemble method for clustering," in *ICANN '01: Proceedings of the International Conference on Artificial Neural Networks.* London, UK: Springer-Verlag, 2001, pp. 217–224. 67

[27] F. J. Duarte, "Optimização da combinação de agrupamentos baseado na acumulação de provas pesadas por índices de validação e com uso de amostragem," Ph.D. dissertation, Universidade de Trás-os-Montes e Alto Douro, 2008 - aguarda defesa. 2, 8, 10, 62, 63, 66, 73, 77, 119

[28] F. J. Duarte, A. L. N. Fred, M. F. C. Rodrigues, and J. Duarte, "Weighted evidence

accumulation clustering using subsampling," in *Sixth International Workshop on Pattern Recognition in Information Systems*, 2006. 64

[29] S. Dudoit and J. Fridlyand, "Bagging to improve the accuracy of a clustering procedure," *Bioinformatics*, vol. 19, no. 9, pp. 1090–1099, 2003. 67, 69

[30] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, 1996, pp. 226–231. 11

[31] X. Z. Fern and C. E. Brodley, "Random projection for high dimensional data clustering: A cluster ensemble approach," 2003, pp. 186–193. 65

[32] X. Fern and C. Brodley, "Solving cluster ensemble problems by bipartite graph partitioning," in *ICML '04: Proceedings of the twenty-first international conference on Machine learning*. New York, NY, USA: ACM, 2004, p. 36. 71

[33] D. H. Fisher, "Knowledge acquisition via incremental conceptual clustering," *Mach. Learn.*, vol. 2, no. 2, pp. 139–172, 1987. 12, 28

[34] R. W. Floyd, "Algorithm 97: Shortest path," *Commun. ACM*, vol. 5, no. 6, p. 345, 1962. 42

[35] A. L. N. Fred, "Finding consistent clusters in data partitions," in *MCS '01: Proceedings of the Second International Workshop on Multiple Classifier Systems*. London, UK: Springer-Verlag, 2001, pp. 309–318. 2, 68, 73, 75, 88, 119

[36] A. L. N. Fred and A. K. Jain, "Combining multiple clusterings using evidence accumulation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 6, pp. 835–850, 2005. 64

[37] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," in *EuroCOLT '95: Proceedings of the Second European Conference on Computational Learning Theory*. London, UK: Springer-Verlag, 1995, pp. 23–37. 61, 65, 73

[38] R. Ge, M. Ester, W. Jin, and I. Davidson, "Constraint-driven clustering," in *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2007, pp. 320–329. 21

[39] B. German, UCI Machine Learning Repository [http://www.ics.uci.edu/ mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science., 1987. 88

[40] S. Guha, R. Rastogi, and K. Shim, "Cure: an efficient clustering algorithm for large databases," in *SIGMOD '98: Proceedings of the 1998 ACM SIGMOD international conference on Management of data*. New York, NY, USA: ACM, 1998, pp. 73–84. 10

[41] S. T. Hadjitodorov, L. I. Kuncheva, and L. P. Todorova, "Moderate diversity for better cluster ensembles," *Inf. Fusion*, vol. 7, no. 3, pp. 264–275, 2006. 64

[42] J. F.-J. Hou, "Clustering with obstacle entities," 1999. [Online]. Available: citeseer.ist.psu.edu/hou99clustering.html 18

[43] A. K. Jain and F. Farrokhnia, "Unsupervised texture segmentation using gabor filters," *Pattern Recogn.*, vol. 24, no. 12, pp. 1167–1186, 1991. 20

[44] P. Jouve and N. Nicoloyannis, "A new method for combining partitions, applications for distributed clustering," in *International Workshop on Paralell and Distributed Machine Learning and Data Mining (ECML/PKDD03)*, 2003, pp. 35–46. 70

[45] P.-E. Jouve and N. Nicoloyannis, "Kerouac : an algorithm for clustering categorical data sets with practical advantages," in *International Workshop on Data Mining for Actionable Knowledge (DMAK'2003, in conjunction with PAKDD03)*, 2003. 70

[46] G. Karypis and V. Kumar, "Metis, a software package for partitioning unstructured graphs, partitioning meshes, and computing fill-reducing orderings of sparse matrices," Department of Computer Science/Army HPC Research Centre, University of Minnesota, Minneapolis, MN, Technical Report, 1997. 71

[47] G. Karypis, E.-H. S. Han, and V. Kumar, "Chameleon: Hierarchical clustering using dynamic modeling," *Computer*, vol. 32, no. 8, pp. 68–75, 1999. 11

[48] G. Karypis and V. Kumar, "A fast and high quality multilevel scheme for partitioning irregular graphs," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 359–392, 1998. 70

[49] L. Kaufmann and P. J. Rousseeuw, "Clustering by means of medoids," In Dodge, Y. (Ed.) Statistical Data Analysis based on the L1 Norm. pp. 405-416. Elsevier/North Holland, Amsterdam, 1987. 9

[50] P. Kellam, X. Liu, N. Martin, C. Orengo, S. Swift, A., and Tucker, "Comparing, contrasting and combining clusters in viral gene expression data," in *Proceedings of the Intelligent Data Analysis in Medicine and Pharmacology Workshop (IDAMAP–2001)*, London, UK, 2001, pp. 56–62. 68

[51] B. King, "Step-wise clustering procedures," *Journal of the American Statistical Association*, no. 69, pp. 86–101, 1963. 10

[52] D. Klein, S. D. Kamvar, and C. D. Manning, "From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering," in *ICML '02: Proceedings of the Nineteenth International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2002, pp. 307–314. 41

[53] T. Kohonen, "The self-organizing map," in *Proceedings of the IEEE*, vol. 78, no. 9, 1990, pp. 1464–1480. 12

[54] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistic Quarterly*, vol. 2, pp. 83–97, 1955. 67

[55] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 2004. 58, 59

[56] N. Littlestone and M. K. Warmuth, "The weighted majority algorithm," *Inf. Comput.*, vol. 108, no. 2, pp. 212–261, 1994. 60

[57] Z. Lu and T. K. Leen, "Penalized probabilistic clustering," *Neural Comput.*, vol. 19, no. 6, pp. 1528–1567, 2007. 47, 49

[58] J. B. Macqueen, "Some methods of classification and analysis of multivariate obser-

vations," in *Proceedings of the Fifth Berkeley Symposium on Mathemtical Statistics and Probability*, 1967, pp. 281–297. 9, 29

[59] C. J. Merz, "Dynamical selection of learning algorithms," in *In*. Springer-Verlag, 1995. 60

[60] R. Mihalcea, "Co-training and self-training for word sense disambiguation," in *In CoNLL-2004*, 2004, pp. 33–40. 14

[61] B. Minaei-Bidgoli, A. Topchy, and W. F. Punch, "Ensembles of partitions via data resampling," in *ITCC '04: Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'04) Volume 2*. Washington, DC, USA: IEEE Computer Society, 2004, p. 188. 64

[62] B. Mirkin, "Reinterpreting the category utility function," *Mach. Learn.*, vol. 45, no. 2, pp. 219–228, 2001. 69

[63] K. Nakai, UCI Machine Learning Repository [http://www.ics.uci.edu/ mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science., 1996. 87

[64] R. M. Neal, "Probabilistic inference using Markov chain Monte Carlo methods," University of Toronto, Tech. Rep. CRG-TR-93-1, 1993. [Online]. Available: citeseer.ist.psu.edu/neal93probabilistic.html 49

[65] R. T. Ng and J. Han, "Efficient and effective clustering methods for spatial data mining," in *VLDB '94: Proceedings of the 20th International Conference on Very Large Data Bases*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1994, pp. 144–155. 9

[66] R. Ng and J. Han, "Clarans: A method for clustering objects for spatial data mining," *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 5, pp. 1003–1016, 2002. 18

[67] K. P. Nigam, "Using unlabeled data to improve text classification," Tech. Rep., 2001. 14

[68] M. A. Oliver and R. Webster, "A geostatistical basis for spatial weighting in multivariate classification," *Mathematical Geology*, vol. 21, pp. 15–21, 1989. 20

[69] D. Opitz and R. Maclin, "Popular ensemble methods: An empirical study," *Journal of Artificial Intelligence Research*, vol. 11, pp. 169–198, 1999. 61

[70] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine*, vol. 2, no. 6, pp. 559–572, 1901. 65

[71] T. Pedersen, "A simple approach to building ensembles of naive bayesian classifiers for word sense disambiguation," in *Proceedings of the first conference on North American chapter of the Association for Computational Linguistics*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000, pp. 63–69. 60

[72] D. Pelleg and D. Baras, "K-means with large and noisy constraint sets," in *The 18th European Conference on Machine Learning*, 2007. 32, 38

[73] J. Quinlan, "Introduction of decision trees," *Machine Learning*, vol. 1, pp. 81–106, 1986. 6

[74] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," pp. 318–362, 1986. 6

[75] D. B. Shmoys, E. Tardos, and K. Aardal, "Approximation algorithms for facility location problems," in *In Proceedings of the 29th Annual ACM Symposium on Theory of Computing*, 1997, pp. 265–274. 22

[76] P. Sneath and R. Sokal, *Numerical Taxonomy*. San Francisco, California: W. H. Freeman, 1973. 10

[77] A. Strehl and J. Ghosh, "Cluster ensembles — a knowledge reuse framework for combining multiple partitions," *J. Mach. Learn. Res.*, vol. 3, pp. 583–617, 2003. 69, 70

[78] J. Theiler and G. Gisler, "A contiguity-enhanced k-means clustering algorithm for unsupervised multispectral image segmentation," 1997. [Online]. Available: citeseer.ist.psu.edu/theiler97contiguityenhanced.html 20

[79] A. Topchy, A. K. Jain, and W. Punch, "Combining multiple weak clusterings," in *ICDM '03: Proceedings of the Third IEEE International Conference on Data Mining*. Washington, DC, USA: IEEE Computer Society, 2003, p. 331. 64, 65, 69

[80] A. Topchy, B. Minaei-Bidgoli, A. K. Jain, and W. F. Punch, "Adaptive clustering ensembles," in *ICPR '04: Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 1*. Washington, DC, USA: IEEE Computer Society, 2004, pp. 272–275. 65

[81] A. P. Topchy, A. K. Jain, and W. F. Punch, "A mixture model for clustering ensembles." in *SDM*, M. W. Berry, U. Dayal, C. Kamath, and D. B. Skillicorn, Eds. SIAM, 2004. 73

[82] A. K. H. Tung, J. Han, V. S. Lakshmanan, and R. T. Ng, "Constraint-based clustering in large databases," *Lecture Notes in Computer Science*, vol. 1973, pp. 405–??, 2001. [Online]. Available: citeseer.ist.psu.edu/tung00constraintbased.html 21

[83] B. V. and F. A., "Constrained clustering problems," 1998. 20

[84] R. Vilalta and Y. Drissi, "A perspective view and survey of meta-learning," *Artif. Intell. Rev.*, vol. 18, no. 2, pp. 77–95, 2002. 60

[85] K. Wagstaff and C. Cardie, "Clustering with instance-level constraints," in *Proceedings of the Seventeenth International Conference on Machine Learning*, 2000, pp. 1103–1110. [Online]. Available: citeseer.ist.psu.edu/wagstaff00clustering.html 2, 28, 119

[86] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl, "Constrained k-means clustering with background knowledge," in *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 577–584. 21, 29

[87] K. L. Wagstaff, "Intelligent clustering with instance-level constraints," Ph.D. dissertation, Ithaca, NY, USA, 2002, chair-Claire Cardie. 18, 19, 23, 24

[88] W. Wang, J. Yang, and R. R. Muntz, "Sting: A statistical information grid approach to spatial data mining," in *VLDB '97: Proceedings of the 23rd International Conference on Very Large Data Bases*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1997, pp. 186–195. 12

[89] J. Ward, "Hierarchical grouping to optimize an objective function," *Journal of the Ame-*

rican Statistical Association, no. 58, pp. 236–244, 1963. 10

[90] W. Wolberg, UCI Machine Learning Repository [http://www.ics.uci.edu/ mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science., 1992. 87

[91] D. H. Wolpert, "Stacked generalization," *Neural Netw.*, vol. 5, no. 2, pp. 241–259, 1992. 60

[92] J. Wu, Z. Zhou, and Z. Chen, "Ensemble of ga based selective neural network ensembles," in *Proceedings of the 8th International Conference on Neural Information Processing*, vol. 3, 2001, pp. 1477–1482. 60

[93] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell, "Distance metric learning, with application to clustering with side-information," in *Advances in Neural Information Processing Systems 15*. MIT Press, 2003, pp. 505–512. 42

[94] S. Yu, "Feature selection and classifier ensembles: A study on hyperspectral remote sensing data," Ph.D. dissertation, Universiteit Antwerpen, 2003. 60

[95] X. Zhu, "Semi-supervised learning literature survey," Computer Sciences, University of Wisconsin-Madison, Tech. Rep. 1530, 2005. 14, 23