



Universidade do Porto
Faculdade de Engenharia

FEUP



Sara Lúcia Ferreira Tavares Pinto

Modelos Probabilísticos em Classificação Não Hierárquica

519.2(043)
PINs/MOD

Abril, 2005

**FACULDADE DE ENGENHARIA DA
UNIVERSIDADE DO PORTO**

Modelos Probabilísticos em Classificação Não Hierárquica

Sara Lúcia Ferreira Tavares Pinto
Licenciada em Matemática — Ramo Educacional pela
Faculdade de Ciências da Universidade do Porto

Dissertação submetida para satisfação parcial dos requisitos do grau de mestre
em
Estatística Aplicada e Modelação

Dissertação realizada sob a supervisão da Professora Doutora Paula Brito
da Faculdade de Economia da Universidade do Porto
e da Professora Doutora Fernanda Sousa
da Faculdade de Engenharia da Universidade do Porto

Porto, Abril de 2005

Agradecimentos

À Professora Doutora Fernanda Sousa e à Professora Doutora Paula Brito, orientadoras desta tese, quero agradecer pela excelência da sua orientação. Muito obrigado pela disponibilidade, interesse e profissionalismo que sempre demonstraram ao longo do desenvolvimento deste trabalho.

À Faculdade de Engenharia da Universidade do Porto e à Direcção do Mestrado agradeço todas as condições que disponibilizaram para a realização deste trabalho.

Às minhas amigas, Fatucha e Filipa, e ao Jorge, quero agradecer pelo companheirismo, apoio e espírito de entre-ajuda.

Aos meus pais, irmã, e em especial, à minha querida avó, um grande obrigado pelo apoio dado ao longo deste período.

Ao Celso, sempre a meu lado, dedico este trabalho.

UNIVERSIDADE DO PORTO
Faculdade de Engenharia
BIBLIOTECA M
N.º 72775
CDU _____
Data 29.11, 2005

Resumo

A dissertação apresentada visa o estudo de métodos de Classificação Não Hierárquica que se inserem na metodologia VL.

A metodologia VL, Validade da Ligação, acenta no uso de medidas de probabilidade para comparar elementos e/ou classes. Os métodos classificatórios não hierárquicos abordados usam como critério de afectação de um elemento a uma classe, a função de distribuição de uma estatística amostral sobre os valores de comparação entre o elemento a classificar e os elementos da classe. Os métodos VL estudados recorrem a estatísticas amostrais de tendência central, nomeadamente, à média aritmética, à média geométrica, à mediana e ao centro, resultando os métodos NHMEAN, NHVGM, NHVMED e NHVC. Simultaneamente, realiza-se um estudo sobre os correspondentes métodos directos, cujo critério de afectação assenta na aplicação directa da estatística sobre os valores de comparação entre elementos.

Em geral, um procedimento classificatório não hierárquico inicia-se com um conjunto de centros que assumem o papel de pólos de agregação. Com base no critério *wise-choice* e considerando a estatística definida em cada um dos métodos, propõem-se novos critérios de inicialização.

O desempenho dos métodos foi avaliado face a conjuntos de dados simulados, com estruturas inicialmente impostas, usando como medida de comparação, entre a partição inicial e a estrutura fornecida, o Índice de Rand Modificado. Os métodos foram também aplicados a conjuntos de dados reais já estudados na literatura.

Abstract

The objective of the present dissertation is to study methods of Non-Hierarchical Clustering that come within the VL methodology.

The VL - Validity of the Link - methodology is based in the use of probability measures to compare elements and/or classes. The considered non-hierarchical clustering methods use as criterion to assign an element to a class the distribution function of a sample statistics of the comparison values between the element to assign and the classe's elements. The studied VL methods use sample central statistics, namely, the arithmetic mean, the geometrical mean, the median and the centre, leading respectively to the methods NHMEAN, NHVGM, NHVMED and NHVC. Simultaneously, a study of the corresponding direct methods, whose assignment criteria is based on the direct application of the statistics of the values of comparison between elements, is performed.

Generally a non-hierarchical clustering process is initialised with a set of centres, that act as aggregation poles. New initialisation criteria are proposed, based on the wise-choice criterion and considering the specific statistics defined in each of the methods.

The performance of the methods is evaluated by considering simulated data with pre-defined structures, and using the Modified Rand Index to compare the initial and the obtained partitions. The methods are also applied to real data sets, previously studied in the literature.

Índice

Introdução	1
1 Medidas de Comparação	5
1.1 Os Dados	5
1.2 Tipos de Variáveis	6
1.3 Função de Comparação entre Elementos	7
1.3.1 Abordagem Geométrica	9
1.3.2 Abordagem Probabilística: Metodologia VL	12
1.4 Comparação entre Classes	16
1.4.1 Função de Comparação entre Classes	16
1.4.2 Comparação entre Classes: Metodologia VL	18
2 Classificação Não Hierárquica	21
2.1 Critérios de Determinação dos Centros Iniciais	23
2.2 Alguns Métodos Centróides	26
2.2.1 Método Forgey	27
2.2.2 Método das k -Médias de MacQueen	28
2.2.3 Método k -Médias de MacQueen: Fusão e Refinamento	32
2.2.4 Método k -Médias de MacQueen: Variante de Wishart	33
2.2.5 Método k -Medóides	35
2.3 Método das Nuvens Dinâmicas	38
2.3.1 Variante dos Centros de Gravidade	40
3 Modelos Probabilísticos em Classificação Não Hierárquica	43
3.1 Métodos VL: Abordagem probabilística	44
3.1.1 Método NHMEAN	46

3.1.2	Método NHVGM	47
3.1.3	Método NHVMED	48
3.1.4	Método NHVC	54
3.2	Métodos Directos: Abordagem Clássica	55
3.2.1	Método NHDMEAN	56
3.2.2	Método NHDGM	56
3.2.3	Método NHDMED	57
3.2.4	Método NHDC	57
3.3	Crítério <i>Wise-Choice</i>	58
3.4	Algoritmo Básico	61
3.5	Algoritmo com ajustamento do número de classes: Fusão e Refinamento	67
3.5.1	Fusão de classes	68
3.5.2	Refinamento de classes	68
4	Avaliação de Partições	71
4.1	Medidas de Homogeneidade de uma Classe	71
4.2	Medidas de Isolamento	74
4.3	Avaliação Global de uma Partição	75
4.4	Comparação de Partições	77
5	Aplicações	85
5.1	Dados Simulados	85
5.1.1	Distribuição Multinormal	86
5.1.2	Resultados	87
5.2	Dados reais	98
5.2.1	Dados das Íris	98
5.2.2	Dados PHYTOS	104
5.2.3	Dados Ruspini	105
5.2.4	Dados de Nutrientes	109
6	Estrutura do Programa	113
6.1	Unidade Principal do Programa	113
6.1.1	Opções de Entrada	114
6.1.2	Opções de Inicialização dos Centros	114
6.1.3	Comparação de Partições: critério de estabilidade	115

6.1.4 Subrotinas	115
7 Notas Finais	
Desenvolvimentos Futuros	119
Bibliografia	121

Introdução

Motivação

A Análise Classificatória de Dados tem como finalidade identificar relações estruturais entre elementos de um conjunto a classificar, conduzindo à formação de grupos que respeitem e evidenciem as características comuns a esses elementos.

O desenvolvimento computacional das últimas décadas proporcionou a implementação de técnicas classificatórias, tornando a Classificação instrumento recorrente de múltiplos domínios (Biologia, Economia, Sociologia, Engenharia,...). De facto, com o crescente armazenamento de grandes conjuntos de dados tornou-se necessário encontrar metodologias capazes de resumir e compreender a informação neles contida. Neste contexto, o recurso a técnicas exploratórias de Análise de Dados Multivariados em geral, e de Classificação em particular, surge como uma solução natural.

A construção de uma partição pode ser vista como um procedimento classificatório que, em geral, consiste em duas etapas. Numa primeira etapa avaliam-se as proximidades entre pares de elementos do conjunto a classificar. O segundo passo baseia-se na aplicação de um critério de agregação entre classes ou de afectação de um elemento a uma classe.

Em Análise Classificatória distinguem-se, fundamentalmente, dois tipos de métodos: os métodos classificatórios hierárquicos que fornecem uma sucessão encaixada de partições, e os métodos classificatórios não hierárquicos que determinam uma partição num número de classes, em geral definido *a priori*.

O trabalho que se apresenta tem como objecto de estudo métodos de Classificação Não Hierárquica que se caracterizam por recorrerem a medidas de probabilidade para avaliarem a ligação entre elementos e/ou classes. Esta metodologia, proposta por Lerman [26], designa-se por metodologia VL, sigla usada para *Validade da Ligação*. O primeiro método VL, denominado por método AVL - *Agregação pela Validade da Ligação*, usa o valor da função de distribuição do máximo das semelhanças entre pares de elementos do conjunto a classificar.

Vários trabalhos têm surgido na continuidade da metodologia iniciada por Lerman. No quadro da Classificação Hierárquica foram propostos os métodos AVM (Agregação pela Validade da Média)(Nicolau [29]), a família paramétrica de métodos VL (Bacelar-Nicolau e Nicolau [4], Nicolau [29]) e os métodos AVMG - *Agregação pela Validade da Média Geométrica*, AVMED - *Agregação pela Validade da Mediana*, e AVC - *Agregação pela Validade do Centro*, que usam como medida de comparação entre classes a função de distribuição de estatísticas de tendência central (Sousa [40]). Em Classificação Não Hierárquica, o uso da metodologia VL deu origem ao método NHMEAN (Nicolau [31], Brito [9]), que usa como critério de afectação de um elemento a uma classe a função de distribuição da média aritmética dos valores de comparação entre o elemento a afectar e os elementos da classe.

Os novos métodos VL de Classificação Não Hierárquica, que nesta dissertação nos propomos estudar, usam critérios de afectação que têm por base o uso de estatísticas amostrais de tendência central, nomeadamente, a média geométrica, a mediana e o centro.

Estrutura do trabalho

Esta dissertação organiza-se em 7 capítulos.

No Capítulo 1, intitulado "Medidas de Comparação", são expostas algumas noções de base importantes para a compreensão do trabalho desenvolvido. Ainda no mesmo capítulo detalha-se a metodologia VL.

No Capítulo 2 é feito um levantamento de alguns métodos clássicos de Classificação

Não Hierárquica (de tipo centróide), como o método das k -médias de MacQueen [28], o método de Forgy [16], o método k -médoides (cf. Kaufman e Rousseeuw [24]) e o método das Nuvens Dinâmicas (Diday [11]). Os métodos aqui expostos requerem a determinação de um conjunto de centros iniciais que assumem o papel de pólos de atracção. No mesmo capítulo é feita uma síntese dos vários critérios de inicialização disponíveis na literatura.

O Capítulo 3 é dedicado à exposição dos novos métodos propostos. Na primeira secção descrevem-se os métodos VL de Classificação que têm por base o uso de estatísticas de tendência central. Na secção seguinte, são apresentados métodos que utilizam uma abordagem directa com base nas mesmas estatísticas, com os quais se pretende avaliar o uso da função de distribuição na função de afectação. A descrição do algoritmo básico, comum aos vários métodos, é efectuada no mesmo capítulo.

Tendo em conta a estatística amostral usada, na Secção 3.3 são sugeridas adaptações ao critério de inicialização *wise-choice*.

A avaliação da homogeneidade e isolamento de classes, bem como a comparação de partições, tem sido tema de estudo por parte de alguns analistas. No Capítulo 4 descrevem-se algumas das medidas disponíveis na literatura, e que serão usadas no trabalho. Simultaneamente, propõem-se medidas de avaliação global das partições que seguem a filosofia VL.

No Capítulo 5 resumem-se os resultados obtidos com os métodos estudados, quando aplicados a conjuntos de dados reais e a conjuntos de dados simulados com uma estrutura imposta.

Remete-se para o Capítulo 6 a exposição da estrutura principal do programa elaborado em FORTRAN PowerStation versão 4.0, e que permitiu obter os resultados numéricos referidos ao longo do trabalho. Aqui, destacam-se alguns detalhes necessários à utilização do programa.

Por último, retomam-se as principais contribuições do trabalho efectuado, apontando vias de desenvolvimento futuro.

Capítulo 1

Medidas de Comparação

Nesta secção são apresentados alguns tópicos basilares na introdução à Análise Classificatória. Introduzem-se noções matemáticas e algumas definições que serão usadas no desenvolvimento deste trabalho.

1.1 Os Dados

De um modo geral, considera-se que os métodos classificatórios operam sobre dois tipos de dados: um quadro de dados iniciais ou um quadro de proximidades. Admite-se que o quadro de dados iniciais é do tipo *indivíduos* $\times$ *variáveis* (figura 1.1), apresentando-se sob a forma de matriz $n \times p$, $X = (x_{ij})$, em que x_{ij} representa o valor da variável Y_j observada no indivíduo I_i , $i = 1, \dots, n$, $j = 1, \dots, p$.

Dada a matriz de dados inicial, alguns métodos classificatórios assumem que as

	Y_1	$\dots$	Y_j	$\dots$	Y_p
I_1			$\vdots$		
$\vdots$			$\vdots$		
I_i	$\dots$	$\dots$	x_{ij}		
$\vdots$					
I_n					

Figura 1.1: Quadro de dados

relações entre todos os pares de elementos podem ser resumidas numa matriz de proximidades, o que pressupõe uma transformação prévia da matriz de dados inicial. Esta matriz de proximidades pode ser calculada para os n indivíduos, dando origem a uma matriz $n \times n$, em que o elemento (i, ℓ) representa a proximidade entre os indivíduos I_i e I_ℓ ; ou para as p variáveis, o que resultará numa matriz $p \times p$, em que o elemento (j, h) representa a proximidade entre as variáveis Y_j e Y_h . Esta matriz de proximidades é geralmente admitida como simétrica e traduz semelhanças ou dissemelhanças entre elementos de acordo com o tipo de função que lhes dá origem.

Neste trabalho, salvo referência em contrário, será designado por E o conjunto a classificar de n indivíduos.

1.2 Tipos de Variáveis

Em geral, a natureza dos dados condiciona o seu tratamento estatístico, pelo que a identificação do tipo de variáveis deve ser uma das primeiras preocupações do analista.

Cada parâmetro escolhido pelo analista para descrever os indivíduos define uma ou mais variáveis. Uma variável define-se por um conjunto O , que se designa por espaço da observação, uma estrutura C sobre O , e uma aplicação Y de E em O , $Y:E \rightarrow O$. Em função da estrutura C , faz-se a distinção entre dois grandes tipos de variáveis: variáveis qualitativas e variáveis quantitativas.

Variáveis qualitativas

O espaço de observação O , em geral é finito. Os elementos de O designam-se por modalidades da variável. Distinguem-se essencialmente os seguintes tipos de variáveis:

- Y é uma **variável qualitativa nominal**, se O não está munido de uma estrutura (ex.: nacionalidade, profissão, ...).

Em particular, se Y assume apenas dois valores, em geral, "0" e "1", diz-se que Y é uma **variável binária**.

- Y é uma **variável qualitativa ordinal**, se O está munido de uma estrutura de ordem total, permitindo a comparação entre duas quaisquer modalidades de Y (ex.: totalmente de acordo, sem opinião, em desacordo).

Variáveis quantitativas

- Y é uma **variável quantitativa** ou **numérica**, se O é um conjunto de números munido de uma estrutura de ordem e operações usuais, como a adição, a multiplicação, ou mesmo o cálculo de médias.

Um quadro de dados pode ser classificado quanto ao tipo de variáveis que o compõem como:

- **Quadro de dados quantitativos:** contém apenas variáveis quantitativas. Neste caso, cada linha do quadro associado é um vector de $\mathbb{R}^p$.
- **Quadro de dados qualitativos ou de modalidades:** contém apenas variáveis qualitativas.
Se todas as variáveis são ordinais (respectivamente, nominais) diz-se um quadro de dados de modalidades ordenadas (respectivamente, não ordenadas).
- **Quadro de dados binários:** as variáveis que constam no quadro de dados são do tipo binário.
- **Quadro de dados heterogéneo:** as variáveis que constam no quadro de dados são de diferentes tipos.

1.3 Função de Comparação entre Elementos

A escolha de uma função de comparação entre pares de elementos está presente na maioria dos métodos de classificação. Trata-se, em geral, de um dos primeiros passos da Análise Classificatória, que visa associar um valor numérico a cada par de elementos, e que mede o seu grau de similitude. Dado E , define-se função de comparação entre elementos de E como uma aplicação $\gamma : E \times E \longrightarrow \mathbb{R}_0^+$.

Uma função de comparação pode ser do tipo:

- **Dissemelhança:** $\gamma(I_i, I_\ell) = d(I_i, I_\ell)$, para dois quaisquer elementos I_i e I_ℓ de E .
Um pequeno valor de $d(I_i, I_\ell)$ significa que I_i e I_ℓ são bastante similares.

- **Semelhança:** $\gamma(I_i, I_\ell) = s(I_i, I_\ell)$, para dois quaisquer elementos I_i e I_ℓ de E .

Um pequeno valor de $s(I_i, I_\ell)$ traduz grandes diferenças entre I_i e I_ℓ .

De seguida, apresentam-se algumas propriedades desejáveis para uma função de comparação quando esta é do tipo dissemelhança.

Seja $d : E \times E \longrightarrow \mathbb{R}_0^+$ uma aplicação e considerem-se as condições:

- i) $d(I_i, I_i) = 0, \forall I_i \in E$.
- ii) $d(I_i, I_\ell) = d(I_\ell, I_i), \forall I_i, I_\ell \in E$ (simetria).

Se d satisfaz i) e ii) diz-se um **índice de dissemelhança**, ou apenas, uma **dissemelhança**.

- iii) $d(I_i, I_\ell) = 0 \Rightarrow I_i = I_\ell$.

Se d é um índice de dissemelhança e satisfaz iii) diz-se um **índice de distância**.

- iv) $d(I_i, I_\ell) \leq d(I_i, I_{i'}) + d(I_{i'}, I_\ell), \forall I_i, I_{i'}, I_\ell \in E$ (desigualdade triangular).

Se d é um índice de distância e satisfaz iv) diz-se uma **métrica** ou **distância**.

- v) $d(I_i, I_\ell) \leq \max\{d(I_i, I_{i'}), d(I_{i'}, I_\ell)\}, \forall I_i, I_{i'}, I_\ell \in E$.

Se d é um índice de distância que verifica v) diz-se uma **distância ultramétrica**.

É óbvio que v) $\Rightarrow$ iv), pelo que toda a distância ultramétrica é uma métrica.

Fazendo uma correspondência para uma função de comparação entre elementos do tipo semelhança, tem-se:

Seja $s : E \times E \longrightarrow \mathbb{R}_0^+$ uma aplicação e considerem-se as condições:

- i) $s(I_i, I_i) = s_{\max}, \forall I_i \in E$.
- ii) $s(I_i, I_\ell) = s(I_\ell, I_i), \forall I_i, I_\ell \in E$.

Se s satisfaz i) e ii) diz-se um **índice de semelhança**, ou apenas, uma **semelhança**.

iii) $s(I_i, I_\ell) = s_{max} \Rightarrow I_i = I_\ell$.

Se s é um índice de semelhança e satisfaz iii) diz-se um **índice de proximidade**.

iv) $s_{max} + s(I_i, I_\ell) \geq s(I_i, I_{i'}) + s(I_{i'}, I_\ell), \forall I_i, I_{i'}, I_\ell \in E$.

Se s é um índice de proximidade e satisfaz iv) diz-se uma **proximidade**.

v) $s(I_i, I_\ell) \geq \min \{s(I_i, I_{i'}), s(I_{i'}, I_\ell)\}, \forall I_i, I_{i'}, I_\ell \in E$.

Se s é um índice de proximidade que verifica v) diz-se uma **proximidade ultramétrica**.

A transformação de uma semelhança $s(I_i, I_\ell)$ numa dissemelhança $d(I_i, I_\ell)$ sobre o mesmo conjunto, pode ser obtida através de uma função decrescente, como

$$d(I_i, I_\ell) = \text{constante} - s(I_i, I_\ell).$$

Usualmente, a constante é o valor máximo que a função s toma, s_{max} . Inversamente, dada uma função de dissemelhança d , pode obter-se uma função de semelhança s , usando a relação

$$s(I_i, I_\ell) = \frac{1}{1 + d(I_i, I_\ell)}.$$

1.3.1 Abordagem Geométrica

Variáveis Numéricas

De seguida, indicam-se alguns exemplos de funções de comparação, do tipo dissemelhança, entre dois indivíduos, I_i e I_ℓ , descritos por p variáveis numéricas.

1. Distâncias de Minkowsky

A família de distâncias de Minkowsky é dada pela fórmula geral

$$d(I_i, I_\ell) = \sqrt[\lambda]{\sum_{j=1}^p |x_{ij} - x_{\ell j}|^\lambda}.$$

Esta distância depende de um parâmetro positivo λ .

Se $\lambda = 2$ tem-se a distância euclideana (ou métrica L_2):

Distância Euclideana

$$d(I_i, I_\ell) = \sqrt{\sum_{j=1}^p (x_{ij} - x_{\ell j})^2}.$$

Para $\lambda = 1$, obtém-se a distância *city block* ou distância de *Manhattan* (métrica L_1):

Distância de *Manhattan* ou distância *city block*

$$d(I_i, I_\ell) = \sum_{j=1}^p |x_{ij} - x_{\ell j}|.$$

Para $\lambda \rightarrow \infty$, tem-se a distância de Chebyshev (métrica L_∞):

Distância de Chebyshev

$$\begin{aligned} d(I_i, I_\ell) &= \lim_{\lambda \rightarrow \infty} \sqrt[\lambda]{\sum_{j=1}^p |x_{ij} - x_{\ell j}|^\lambda} \\ &= \sup_{1 \leq j \leq p} |x_{ij} - x_{\ell j}|. \end{aligned}$$

Distância de Minkowsky Ponderada

Pretendendo reduzir o efeito resultante do uso de diferentes escalas de medição e de diferentes variâncias é possível aplicar pesos às métricas de Minkowsky. Neste caso, tem-se a métrica de Minkowsky ponderada

$$d(I_i, I_\ell) = \sqrt[\lambda]{\sum_{j=1}^p w(j) |x_{ij} - x_{\ell j}|^\lambda}$$

onde $w(j)$ representa o peso da j -ésima variável.

2. Distância Canberra

$$d(I_i, I_\ell) = \sum_{j=1}^p \frac{|x_{ij} - x_{\ell j}|}{|x_{ij}| + |x_{\ell j}|}.$$

Para os valores de j tais que $x_{ij} = x_{\ell j} = 0$ considera-se que a respectiva parcela é nula.

3. Coeficiente de Sokal e Sneath

$$d(I_i, I_\ell) = \sqrt{\frac{1}{p} \sum_{j=1}^p \left(\frac{x_{ij} - x_{\ell j}}{x_{ij} + x_{\ell j}} \right)^2}.$$

Para os valores de j tais que x_{ij} é simétrico de $x_{\ell j}$ considera-se que a respectiva parcela é nula.

Seguem-se dois coeficientes que medem a semelhança entre variáveis, definidos pelo cosseno do ângulo entre dois vectores, sendo estes medidos respectivamente, a partir da origem, e do centro de gravidade do conjunto de pontos que constituem a amostra.

1. Coeficiente de Separação Angular

$$s(Y_j, Y_h) = \frac{\sum_{i=1}^n x_{ij} x_{ih}}{\sqrt{\sum_{i=1}^n x_{ij}^2 \sum_{i=1}^n x_{ih}^2}}.$$

2. Coeficiente de Correlação

$$s(Y_j, Y_h) = \frac{\sum_{i=1}^n (x_{ij} - \bar{x}_{.j})(x_{ih} - \bar{x}_{.h})}{\sqrt{\sum_{i=1}^n (x_{ij} - \bar{x}_{.j})^2 \sum_{i=1}^n (x_{ih} - \bar{x}_{.h})^2}}$$

$$\text{onde } \bar{x}_{.j} = \frac{1}{n} \sum_{i=1}^n x_{ij} \text{ e } \bar{x}_{.h} = \frac{1}{n} \sum_{i=1}^n x_{ih}.$$

Variáveis Binárias

Dados dois indivíduos I_i e I_ℓ de E , descritos por p atributos binários, é possível representá-los por dois vectores do espaço $\{0, 1\}^p$, onde cada componente do vector assume o valor 1 se o atributo está presente no indivíduo, e assume o valor 0 em caso de ausência de atributo. Relativamente a um atributo, diz-se que I_i e I_ℓ têm uma concordância (ou associação) positiva (respectivamente, negativa) se o atributo está presente (respectivamente, ausente) nos dois indivíduos. Se o atributo está presente

em I_i e ausente em I_l (ou vice-versa) tem-se uma discordância. Representando por a o número de concordâncias positivas, d o número de concordâncias negativas, b e c , o número de discordâncias de cada tipo, então toda a informação necessária para medir a proximidade entre I_i e I_l está contida na tabela de contingência representada na tabela 1.1.

$\begin{smallmatrix} I_l \\ \backslash I_i \end{smallmatrix}$	1	0	
1	a	b	$a + b = n_{I_i}$
0	c	d	$p - n_{I_i}$
	$a + c = n_{I_l}$	$p - n_{I_l}$	$p = a + b + c + d$

Tabela 1.1: Tabela de Contingência 2×2

A tabela 1.2 lista alguns dos índices de comparação mais usuais para dados do tipo binário.

	$s(I_i, I_l)$	$d(I_i, I_l)$
Índice de Russel e Rao (1940)	$\frac{a}{p}$	$\frac{b + c + d}{p}$
Índice de Sokal e Michener (1958)	$\frac{a + d}{p}$	$\frac{b + c}{p}$
Índice de Jacard (1901)	$\frac{a}{a + b + c}$	$\frac{b + c}{a + b + c}$
Índice de Kulczynski (1927)	$\frac{a/(a + b) + a/(a + c)}{2}$	$1 - \frac{a/(a + b) + a/(a + c)}{2}$
Índice de Sokal e Sneath	$\frac{a}{a + 2(b + c)}$	$\frac{2(b + c)}{a + 2(b + c)}$
Índice de Dice (1945)	$\frac{2a}{2a + b + c}$	$\frac{b + c}{2a + b + c}$
Índice de Rogers e Tanimoto (1960)	$\frac{a + d}{a + d + 2(b + c)}$	$\frac{2(b + c)}{a + d + 2(b + c)}$
Índice de Ochiai (1957)	$\frac{a}{\sqrt{(a + b)(a + c)}}$	$1 - \frac{a}{\sqrt{(a + b)(a + c)}}$

Tabela 1.2: Índices de comparação para dados binários

1.3.2 Abordagem Probabilística: Metodologia VL

As medidas de comparação entre elementos atrás enumeradas, derivam de uma concepção geométrica. Uma abordagem diferente foi introduzida por Lerman [26], e

desenvolvida por Bacelar-Nicolau [3] e Nicolau [29]. Esta abordagem assenta numa concepção probabilística, dando origem ao índice de semelhança VL, sigla usada para a expressão “Validade da Ligação” ou “Verosimilhança da Ligação”. A obtenção de índices deste tipo contempla os seguintes passos:

1. Escolhe-se um índice de semelhança s , apropriado ao tipo de dados;
2. Consideram-se os valores da medida de comparação entre elementos $s(I_i, I_\ell)$ realizações de uma variável aleatória S , cuja função de distribuição se determina relativamente a uma hipótese de referência, que neste trabalho será a de independência.

Dados $I_i, I_\ell \in E$, a semelhança entre I_i e I_ℓ , $\gamma(I_i, I_\ell)$, é agora obtida pelo valor da função de distribuição de S no ponto $s(I_i, I_\ell)$. Esquematizando:

$$(I_i, I_\ell) \longrightarrow s(I_i, I_\ell) \xrightarrow{f.d.} \gamma(I_i, I_\ell) = P(S \leq s(I_i, I_\ell)) = F_S(s(I_i, I_\ell)).$$

A aplicação da função de distribuição a uma variável aleatória resulta numa variável aleatória uniformemente distribuída em $[0, 1]$ (propriedade uniformizante). Logo, os valores do conjunto $\{\gamma(I_i, I_\ell) : I_i, I_\ell \in E\}$ não são mais do que realizações de uma variável aleatória de lei uniforme em $[0, 1]$.

Deduzida a lei de probabilidade de S , calcula-se o valor esperado de S , $E(S)$, e a sua variância, $Var(S)$. Centrando e reduzindo S , obtém-se a variável aleatória

$$S^* = \frac{S - E(S)}{\sqrt{Var(S)}}$$

que é assintoticamente normal reduzida, isto é,

$$S^* \xrightarrow{L} N(0, 1)^1.$$

Tem-se então que:

$$\gamma(I_i, I_\ell) = P(S \leq s(I_i, I_\ell)) = P(S^* \leq s^*(I_i, I_\ell)) \cong \Phi(s^*(I_i, I_\ell))$$

onde Φ representa a função de distribuição normal reduzida, $N(0, 1)$. Designa-se $\gamma(I_i, I_\ell)$ por **semelhança VL**, entre o par de elementos I_i e I_ℓ .

¹Convergência em lei.

Repare-se que a variável aleatória S , embora discreta, quando considerada num conjunto suficientemente grande goza da propriedade uniformizante. Este facto facilita o tratamento estatístico da semelhança VL, dado o conhecimento do comportamento distribucional de amostras de variáveis uniformes. Neste trabalho terão particular interesse as leis da média aritmética, média geométrica, da mediana e do centro de variáveis aleatórias independentes e uniformemente distribuídas em $[0, 1]$.

Escolhendo um índice de dissemelhança d e tomando os valores de d como concretizações de uma variável aleatória D , então neste caso, a validade da ligação de dois indivíduos I_i e I_l , $\gamma(I_i, I_l)$, é dada por

$$(I_i, I_l) \xrightarrow{f.d.} d(I_i, I_l) \xrightarrow{f.d.} \gamma(I_i, I_l) = P(D \geq d(I_i, I_l)) = 1 - P(D \leq d(I_i, I_l)) = 1 - F_D(d(I_i, I_l)).$$

A aplicação da metodologia VL a atributos binários é brevemente descrita no parágrafo seguinte, uma vez que os métodos desenvolvidos serão também aplicados a dados deste tipo. Para uma análise mais detalhada consulte-se Bacelar-Nicolau [3].

Observe-se novamente a tabela de contingência 2×2 representada na tabela 1.1 associada aos indivíduos I_i e I_l . Em Classificação, na análise de uma tabela de contingência 2×2 , uma de duas hipóteses pode ocorrer: margens livres ou margens fixas.²

Na situação de margens fixas, $a + b = n_{I_i}$ e $a + c = n_{I_l}$ são valores fixos, o que acrescido ao facto de $a + b + c + d = p$, deixa a tabela com apenas um grau de liberdade na variação dos seus elementos a , b , c e d . Logo, a variável aleatória S definida como o número de concordâncias positivas cuja realização da tabela 1.1 é a , é uma variável hipergeométrica, isto é,

$$S \sim H(p, n_{I_i}, \frac{n_{I_l}}{p}) \text{ ou } S \sim H(p, n_{I_l}, \frac{n_{I_i}}{p}).$$

Logo, a probabilidade de S assumir o valor a é dada por

²A hipótese de uma margem fixa e outra margem livre não é aqui considerada uma vez que, atendendo à simetria de posição entre dois quaisquer indivíduos e à necessidade de reter todos os emparelhamentos, esta hipótese não se enquadra no âmbito do trabalho.

$$P(S = a) = \frac{\binom{n_{I_t}}{a} \binom{p - n_{I_t}}{n_{I_i} - a}}{\binom{p}{n_{I_i}}} = \frac{\binom{n_{I_i}}{a} \binom{p - n_{I_i}}{n_{I_t} - a}}{\binom{p}{n_{I_t}}}.$$

Se $S \sim H(p, n_{I_i}, \frac{n_{I_t}}{p})$, então S terá valor médio

$$E(S) = \frac{n_{I_i} n_{I_t}}{p}$$

e variância

$$Var(S) = \frac{n_{I_i} n_{I_t} (p - n_{I_i})(p - n_{I_t})}{p^2(p - 1)}.$$

Centrando e reduzindo S , obtém-se $S^* = \frac{S - E(S)}{\sqrt{Var(S)}}$, cuja concretização em função de a, d, b e c é dada por:

$$s^* = \frac{ad - bc}{\sqrt{(a + b)(a + c)(b + d)(c + d)}} \times \sqrt{p - 1}.$$

Tem-se que S^* é assintoticamente normal reduzida, logo

$$\gamma(I_i, I_t) = P(S^* \leq s^*(I_i, I_t)) \cong \Phi(s^*(I_i, I_t)).$$

Para o caso de margens livres, a dedução da semelhança VL é efectuada de modo análogo. Considera-se o vector (a, b, c, d) como realização de uma variável aleatória multinomial, então a é a realização de uma variável aleatória binomial S' de parâmetros p e p_{11} , $S' \sim Bi(p, p_{11})$, sendo p_{11} a probabilidade de uma concordância, a qual é estimada por $p_{11} = \frac{n_{I_i} n_{I_t}}{p^2}$, na hipótese de independência entre indivíduos. Decorre de

$$E(S') = \frac{n_{I_i} n_{I_t}}{p}$$

e

$$Var(S') = \frac{n_{I_i} n_{I_t}}{p} \left(1 - \frac{n_{I_i} n_{I_t}}{p^2} \right)$$

que centrando e reduzindo a variável aleatória S' , obtém-se

$$S'^* = \frac{S' - E(S')}{\sqrt{Var(S')}}.$$

tendo como realização

$$s'^* = \frac{(ad - bc)\sqrt{p}}{\sqrt{(a+b)(a+c)[p^2 - (a+b)(a+c)]}}. \quad (1.3.1)$$

S'^* segue assintoticamente uma distribuição normal reduzida, o que na prática permite tomar $\gamma(I_i, I_\ell) = P(S'^* \leq s'^*(I_i, I_\ell)) \cong \Phi(s'^*(I_i, I_\ell))$.

Pretendendo-se determinar um índice de comparação entre variáveis, perante um quadro de dados binários, tem-se por base uma tabela de contingência como a que figura na tabela 1.1, substituindo p por n .

Em todo o trabalho, considera-se que os elementos a classificar são indivíduos. A mesma metodologia pode ser aplicada a variáveis, usando um índice de comparação adequado.

1.4 Comparação entre Classes

1.4.1 Função de Comparação entre Classes

Um procedimento classificatório tem, em geral, como ponto de partida a comparação entre elementos, dando assim origem a um conjunto de valores que medem a sua proximidade. Sobre este conjunto de valores recairá a decisão de agrupar elementos em classes com base num determinado critério. Alguns métodos visam a obtenção directa de uma partição de um conjunto a classificar em k classes, os quais se inserem no quadro da Classificação Não Hierárquica. Em Classificação Hierárquica, os métodos dão origem a uma sucessão de partições encaixadas. Neste caso, a comparação entre classes surge como passo integrante do procedimento classificatório. Importa, então, definir função de comparação entre classes.

Define-se **função de comparação entre partes** de um conjunto finito E como sendo uma aplicação $\Gamma : P(E) \times P(E) \longrightarrow \mathbb{R}_0^+$, onde $P(E)$ representa o conjunto de partes não vazias de E . Dadas duas partes disjuntas, A e B , de elementos de E , a função de comparação entre A e B representa-se por $\Gamma(A, B)$.

Considerem-se A e B , classes não vazias de E , de cardinais n_A e n_B , respectivamente, e γ uma função de comparação entre elementos do tipo dissemelhança. Seguem-

se algumas funções de comparação entre classes propostas na literatura.

I. Ligação Mínima (*Single Linkage*)

$$\Gamma(A, B) = \min\{\gamma(a, b) : (a, b) \in A \times B\}$$

A proximidade entre as duas classes é avaliada pelo mínimo das dissemelhanças entre pares de elementos das duas classes.

II. Ligação Máxima (*Complete Linkage*)

$$\Gamma(A, B) = \max\{\gamma(a, b) : (a, b) \in A \times B\}$$

Na Ligação Máxima usa-se a estatística máximo das dissemelhanças entre pares de elementos das duas classes.

III. Ligação Média

$$\Gamma(A, B) = \sum_{(a,b) \in A \times B} \frac{\gamma(a, b)}{n_A n_B}$$

A ligação entre duas classes A e B é dada pela média aritmética das dissemelhanças entre os pares de elementos pertencentes a $A \times B$.

Considerando, alternativamente, γ uma função de comparação do tipo semelhança, a Ligação Mínima é dada por $\Gamma(A, B) = \max\{\gamma(a, b) : (a, b) \in A \times B\}$ e a Ligação Máxima é medida por $\Gamma(A, B) = \min\{\gamma(a, b) : (a, b) \in A \times B\}$. Obviamente, não há alterações na Ligação Média.

As funções de comparação apresentadas têm como ponto de partida os valores das funções de comparação entre pares de elementos. As medidas que de seguida se enumeram operam sobre a matriz de dados inicial, quando constituída apenas por variáveis quantitativas.

IV. Centróide

O método centróide baseia-se na distância entre os centros de gravidade das duas classes a comparar:

$$\Gamma(A, B) = d(g_A, g_B)$$

sendo $g_A = \frac{1}{n_A} \left(\sum_{I_i \in A} x_{i1}, \dots, \sum_{I_i \in A} x_{ip} \right)$, $g_B = \frac{1}{n_B} \left(\sum_{I_i \in B} x_{i1}, \dots, \sum_{I_i \in B} x_{ip} \right)$ e d a distância euclideana.

V. Ward

O método de Ward baseia-se no conceito de inércia de uma classe. Dadas as classes A e B , considerem-se as quantidades

$$SQD_A = \sum_{I_i \in A} d^2(I_i, g_A) \text{ e } SQD_B = \sum_{I_i \in B} d^2(I_i, g_B)$$

sendo $g_A = \frac{1}{n_A} \left(\sum_{I_i \in A} x_{i1}, \dots, \sum_{I_i \in A} x_{ip} \right)$ e $g_B = \frac{1}{n_B} \left(\sum_{I_i \in B} x_{i1}, \dots, \sum_{I_i \in B} x_{ip} \right)$ e d^2 o quadrado da distância euclideana.

Neste caso,

$$\Gamma(A, B) = SQD_{A \cup B} - (SQD_A + SQD_B),$$

sendo

$$SQD_{A \cup B} = \sum_{I_i \in A \cup B} d^2(I_i, g_{A \cup B})$$

$$\text{e } g_{A \cup B} = \frac{1}{n_A + n_B} \left(\sum_{I_i \in A \cup B} x_{i1}, \dots, \sum_{I_i \in A \cup B} x_{ip} \right).$$

1.4.2 Comparação entre Classes: Metodologia VL

Na Secção 1.3.2 expôs-se a metodologia de uma abordagem probabilística na definição de um índice de comparação entre elementos. A metodologia VL pode ser estendida à definição da função de comparação entre classes.

Dado o conjunto E a classificar, de cardinal n , e escolhida uma função de comparação entre elementos do tipo semelhança (ou dissemelhança), interessa apenas reter os $\frac{n(n-1)}{2}$ pares de elementos distintos para definir a matriz dos valores da função de comparação associada ao conjunto E . Assim, dadas duas classes, A e B , de cardinais n_A e n_B , admite-se que toda a informação útil para avaliar a sua proximidade está contida no conjunto de semelhanças VL $\{\gamma(a, b) : (a, b) \in A \times B\}$, que se considera

uma amostra de $\frac{n_A n_B (n_A n_B - 1)}{2}$ variáveis aleatórias independentes e identicamente distribuídas em $[0, 1]$. A função de comparação entre classes é dada pela função de distribuição de uma estatística U , aplicada sobre o conjunto de semelhanças VL. O uso de diferentes estatísticas origina obviamente diferentes medidas de comparação.

De seguida, indicam-se os primeiros métodos de Classificação Hierárquica do conjunto de métodos VL. Considera-se γ uma função do tipo semelhança.

I. Método AVL (Agregação pela Validade da Ligação)

Para função de comparação entre partes, Lerman [26] propôs o uso da função de distribuição da estatística máximo das semelhanças entre as classes, isto é,

$$\Gamma_{AVL}(A, B) = P(U_{Max} \leq u_{Max}(A, B)) = (u_{Max}(A, B))^{n_A n_B}$$

para $u_{Max} = \max \{\gamma(a, b) : (a, b) \in A \times B\}$, concretização da estatística U_{Max} .

II. Método AVD (Agregação pela Validade do Diâmetro)

Para função de comparação entre partes, o método AVD, proposto por Nicolau [29], baseia-se na menor das ligações entre duas classes

$$\Gamma_{AVD}(A, B) = P(U_D \leq u_D(A, B)) = 1 - (1 - u_D(A, B))^{n_A n_B}$$

para $u_D = \min \{\gamma(a, b) : (a, b) \in A \times B\}$, realização da estatística U_D .

III. Método AVM (Agregação pela Validade da Ligação da Média)

Nicolau [29] define a semelhança entre duas classes A e B , recorrendo ao valor da função de distribuição da estatística média aritmética das semelhanças dos pares de elementos de $A \times B$. A função de comparação entre duas classes, A e B , assume a seguinte forma:

$$\begin{aligned} \Gamma_{AVM}(A, B) &= P(U_M \leq u_M(A, B)) \\ &= \frac{1}{(n_A n_B)!} \sum_{j=0}^{n_A n_B u_M(A, B)} \binom{n_A n_B}{j} (n_A n_B u_M(A, B) - j)^{n_A n_B} \end{aligned}$$

sendo $u_M(A, B) = \frac{\sum_{(a, b) \in A \times B} \gamma(a, b)}{n_A n_B}$ a concretização da estatística U_M .

Com base na mesma abordagem, Nicolau [29] e Bacelar-Nicolau e Nicolau [4] propuseram uma família paramétrica de métodos VL cuja função de comparação entre classes é dada por

$$\Gamma(A, B) = P(U_{Max} \leq u_{Max}(A, B)) = (u_{Max}(A, B))^{1+\varepsilon[(n_A n_B)^\xi - 1]}.$$

Com esta família pretende-se controlar o efeito de cardinalidade. Destaca-se o método AVB, que se obtém para $\varepsilon = 1$ e $\xi = 0.5$.

No Capítulo 3 podem ser vistos outros métodos do tipo VL, que têm por base outras estatísticas de tendência central.

Capítulo 2

Classificação Não Hierárquica

Neste capítulo referem-se alguns métodos de Classificação Não Hierárquica. Os métodos aqui abordados apresentam como principal vantagem a facilidade com que são aplicados a matrizes de dados muito grandes, uma vez que não é necessário calcular e armazenar uma nova matriz de semelhanças ou dissemelhanças em cada passo do algoritmo. Adicionalmente, os métodos não hierárquicos são capazes de reagrupar os sujeitos numa classe diferente daquela em que estes foram inicialmente incluídos, o que não acontece com os métodos de classificação hierárquica onde a inclusão de um sujeito numa determinada classe é definitiva.

Dado um conjunto de indivíduos, o problema que se coloca consiste em construir uma partição cujas classes satisfaçam critérios de coesão interna e isolamento externo. A discordância entre a estrutura do conjunto de dados e a partição obtida é medida por uma função, a qual se designa por "função erro", usualmente associada à soma dos quadrados dos desvios dos indivíduos de cada classe em relação ao respectivo centro de gravidade. A solução "óptima" será a que minimiza esta "função erro". Observe-se, no entanto, que o número de partições de um conjunto de n elementos em k classes, é dado pelo número de Stirling de 2ª espécie, S_n^k :

$$S_n^k = \frac{1}{k!} \sum_{j=1}^k (-1)^{k-j} \binom{k}{j} j^n. \quad (2.0.1)$$

À medida que o valor de n aumenta, rapidamente se torna impraticável o cálculo da função erro para todas as partições possíveis. Por exemplo, para $n = 18$ e $k = 5$, $S_{18}^5 \simeq 2.9 \times 10^{10}$. É então necessário recorrer a métodos de classificação que forneçam

óptimos locais que poderão coincidir, ou não, com o óptimo global.

De acordo com as características dos dados ou da "função erro" considerada existem métodos que permitem a obtenção de uma solução óptima global. É o caso do algoritmo de Fisher [14] que se aplica apenas em conjuntos de indivíduos caracterizados por uma única variável quantitativa ou qualitativa ordinal. Neste algoritmo os indivíduos são ordenados e impõe-se que as classes sejam consecutivas. A "função erro" aditiva nas classes permite a redução considerável do número de partições a considerar.

De um modo geral, os métodos não hierárquicos compreendem os seguintes passos:

1. escolha de uma partição inicial ou de um conjunto de pólos de agregação à volta dos quais as classes se formarão;
2. agregação dos indivíduos às classes.

Diferentes técnicas estão na base destes dois passos, originando métodos de Classificação Não Hierárquica distintos. Os métodos descritos neste capítulo inserem-se no conjunto de métodos que se baseiam num processo iterativo, construindo passo a passo a partição solução. De entre os métodos iterativos, é feita a seguinte distinção:

- **Métodos Centróides:** as classificações são determinadas a partir de um conjunto de pontos, representantes das classes, que funcionam como pólos de atracção, sendo os elementos a classificar (re)afectados em função da sua distância aos centros.
- **Métodos de Transferências:** dada uma partição inicial, procuram uma partição solução iterativamente, transferindo em cada etapa um único elemento de uma classe para outra; a transferência a efectuar é escolhida de modo a corresponder ao acréscimo máximo do critério a maximizar.

Neste capítulo são abordados alguns métodos do tipo centróide mais usuais em Classificação Não Hierárquica, tais como o método de Forgy [16], o método das k -Médias de MacQueen [28], o método k -Medóides (cf. Kaufman e Rousseeuw [24]) e o método das Nuvens Dinâmicas (Diday [11]). Outros métodos de Classificação Não Hierárquica, como o Algoritmo do Líder (cf. Hartigan [20]), o método dos Pólos de Atracção de Lerman [27], do tipo centróide, o método das Partições Centrais de Régnier [35] e o

método de Beale [6], que são métodos de transferências, podem ser vistos em Brito [9].

A secção seguinte aborda um dos problemas subjacentes aos métodos de Classificação Não Hierárquica de tipo centróide: a determinação dos centros iniciais.

2.1 Critérios de Determinação dos Centros Iniciais

Os métodos estudados neste trabalho exigem a determinação de um conjunto inicial de pontos que assumem o papel de pólos de atracção à volta dos quais as classes se formarão.

Alguns métodos de Classificação Não Hierárquica iniciam, não com um conjunto de pólos de agregação, mas sim com uma partição inicial do conjunto que será submetido a uma classificação.

Seguem-se algumas técnicas que permitem determinar k centros iniciais, elementos do conjunto a classificar, ou não.

1. Método de MacQueen [28]

Escolhem-se para centros iniciais os primeiros k elementos de E .

Se a ordem dos dados, de certa forma não influencia o resultado final, então este método é o mais económico e simples.

2. Numeram-se os dados de 1 até n . Tomam-se para centros iniciais os elementos de ordem $\frac{n}{k}, \frac{2n}{k}, \frac{3n}{k}, \dots, \frac{(k-1)n}{k}$ e n (no caso destes valores não serem inteiros consideram-se os valores característicos).

Este método é quase tão simples como o anterior, no entanto, prima por ultrapassar uma eventual tendência para agregar os elementos segundo a ordem de recolha.

3. Subjectivamente, escolhem-se k elementos do conjunto a analisar.

Este método pressupõe que o analista possua, *a priori*, alguma informação sobre o conjunto de dados.

4. Método de MacRae (cf. Brito [9])

Numeram-se os indivíduos de 1 até n e escolhem-se aleatoriamente k números. Este método sugere, eventualmente, a utilização de um gerador de números aleatórios uniformemente distribuídos em $\{1, \dots, n\}$.

5. Geração de k pontos (cf. Anderberg [1])

Sendo o conjunto E descrito por p variáveis, geram-se k pontos de p coordenadas, em que cada coordenada é um valor escolhido aleatoriamente entre os valores observados para essa variável.

6. Método de Forgy [16]

Divide-se a amostra em k classes disjuntas e calcula-se o respectivo centro de gravidade, que se toma para pólo de agregação.

7. Método de Ball e Hall [5]

Ball e Hall sugerem para primeiro pólo de agregação o vector centro de gravidade da amostra. Os restantes centros obtêm-se calculando a distância dos elementos, segundo a ordem de recolha, a todos os centros já calculados. Escolhe-se para centro o primeiro elemento que se encontra a uma distância superior a um valor previamente fixado.

Este processo iterativo é finalizado quando os k centros forem obtidos ou se esgotar todo o conjunto de dados.

8. Escolha dos k pontos mais afastados na amostra e mais afastados entre si (cf. Brito [9])

Para cada indivíduo I_i de E calcula-se o valor médio das semelhanças aos restantes indivíduos, $S(I_i)$

$$S(I_i) = \frac{1}{n-1} \sum_{I_l \in E \setminus \{I_i\}} s(I_i, I_l). \quad (2.1.1)$$

O primeiro centro é o ponto que minimiza (2.1.1),

$$c_1 \equiv I_l \in E : S(I_l) = \min\{S(I_i) : I_i \in E\}. \quad (2.1.2)$$

Os restantes centros c_j , com $j = 2, \dots, k$, deverão satisfazer a condição adicional de se encontrarem o mais afastados possível dos centros já calculados. Assim,

$$\begin{aligned} c_j &\equiv I_l \in E : S(I_l) \times \max_{1 \leq m \leq j-1} s(I_l, c_m) = \\ &= \min\{S(I_i) \times \max_{1 \leq m \leq j-1} s(I_i, c_m) : I_i \in E \setminus \{c_1, \dots, c_{j-1}\}\} \end{aligned} \quad (2.1.3)$$

Na Secção 3.3 propõem-se novos critérios de inicialização com base neste critério.

9. Método de Astrahan (cf. Anderberg [1])

Pretende-se que a escolha dos pólos de atracção seja a mais abrangente possível, na medida em que cada elemento esteja relativamente perto de algum pólo de atracção e estes estejam o mais separados possível uns dos outros. Neste sentido, Astrahan propõe o seguinte processo:

- a) Para cada indivíduo I_i calcula-se a sua *densidade*, den_{I_i} , que se define como o número de indivíduos que se encontram a uma distância inferior a Δ_1 , valor previamente fixado, isto é,

$$den_{I_i} = \text{card}\{I_\ell \in E : d(I_i, I_\ell) < \Delta_1\}.$$

- b) Ordenam-se os elementos por *densidade* e toma-se para primeiro centro, c_1 , o indivíduo que tem maior *densidade*:

$$c_1 \equiv I_\ell \in E : den_{I_\ell} = \max\{den_{I_i} : I_i \in E\}.$$

- c) Na escolha dos restantes centros, pretende-se que a *densidade* diminua e que estejam bem separados dos centros já escolhidos. Para tal, considera-se que o próximo centro c_j , com $j = 2, \dots, k$, deve estar a uma distância Δ_j , superior a Δ_{j-1} , do anterior pólo de agregação c_{j-1} , ou seja, para as distâncias $\Delta_j > \Delta_m$, $1 \leq m < j \leq k$,

$$\begin{aligned} c_j &\equiv I_\ell \in E : den_{I_\ell} = \\ &= \max\{den_{I_i} : I_i \in E \setminus \{c_1, \dots, c_{j-1}\}\} \wedge d(I_\ell, c_m) > \Delta_j, \quad 1 \leq m \leq j-1. \end{aligned}$$

- d) A escolha dos centros é feita até que cada elemento tenha *densidade* nula ou se obtenha o número de centros pretendidos.

A escolha de Δ_1 e restantes distâncias é uma das questões problemáticas que se coloca neste método. A escolha deve ser cuidadosa, pois se Δ_1 tomar valores muito pequenos obter-se-ão indivíduos isolados, conduzindo a uma densidade nula. Pelo contrário, se Δ_1 assumir valores elevados, obter-se-á um reduzido número de pólos.

10. Critério de Lerman

O conjunto de centros iniciais proposto por Lerman [26] baseia-se na maximização da variância das semelhanças.

O primeiro centro c_1 é o elemento do conjunto que maximiza a variância das semelhanças,

$$c_1 \equiv I_t \in E : Var(I_t) = \max\{Var(I_i) : I_i \in E\}$$

sendo

$$Var(I_i) = \frac{1}{n-1} \sum_{I_t \in E \setminus \{I_i\}} (s(I_i, I_t) - S(I_i))^2$$

e

$$S(I_i) = \frac{1}{n-1} \sum_{I_t \in E \setminus \{I_i\}} s(I_i, I_t).$$

Os restantes centros c_j devem estar o mais afastados possíveis dos centros já seleccionados. Assim, para $2 \leq j \leq k$,

$$\begin{aligned} c_j &\equiv I_t \in E : \min_{1 \leq m \leq j-1} \left(\frac{Var(I_t)}{s(c_m, I_t)} \right)^2 = \\ &= \max \left\{ \min_{1 \leq m \leq j-1} \left(\frac{Var(I_i)}{s(c_m, I_i)} \right)^2 : I_i \in E \setminus \{c_1, \dots, c_{j-1}\} \right\} \end{aligned}$$

11. Método *Wise-Choice* (Nicolau [31])

O critério *wise-choice* assenta no princípio de se tomar para centros iniciais elementos do conjunto, bem separados e centrados em vizinhanças densas.

Dado o bom desempenho que este último critério revelou em trabalhos anteriores (Brito [9], Nicolau e Brito [33]), será aqui dada especial atenção ao critério *wise-choice*. Uma análise mais cuidada é realizada na Secção 3.3, onde se descreve o critério e se propõem alterações sobre este, dando assim origem a novos critérios de inicialização.

2.2 Alguns Métodos Centróides

Nos parágrafos seguintes descrevem-se sucintamente um conjunto de métodos do tipo centróide: o método de Forgy [16], o método das k -médias de MacQueen [28] e o método k -medóides introduzido por Vinod (cf. Kaufman e Rousseeuw [24]), e quando possível apresentam-se variantes dos métodos originais. Estes constituem um conjunto de métodos do tipo centróide bastante utilizados.

Para os métodos de Forgy, de MacQueen e suas variantes, assume-se como medida de comparação entre elementos a distância euclidiana.

2.2.1 Método Forgy

Forgy [16] sugere um algoritmo que compreende os seguintes passos:

1. Selecção de k elementos de E , que se tomam para centros iniciais;
2. Afectação dos restantes $n - k$ elementos ao centro mais próximo, mantendo fixos os centros;
3. Cálculo do novo centro de cada classe, que será o centro de gravidade.

Neste processo iterativo são repetidos os passos 2 e 3, até convergência do método que, usualmente, se considera a obtenção da mesma partição em duas iterações consecutivas.

O método de Forgy aplicado a um conjunto de n elementos para obter k classes, requer em cada iteração o cálculo de nk distâncias e $n(k - 1)$ comparações entre as mesmas. Se k for muito inferior a n e o número de iterações para a convergência for pequeno, é possível analisar os diferentes conjuntos de classes para diferentes valores de k , não se tornando ainda assim um processo computacionalmente dispendioso.

Variante de Jancey

Jancey [22] elabora um algoritmo semelhante ao anterior, no entanto, propõe uma alteração no passo 3. No método de Forgy, o conjunto dos centros das classes são os centros de gravidade. Jancey sugere que em cada iteração o novo centro da classe seja obtido "simetrizando" (ou reflectindo) o anterior centro da classe em relação ao novo centro de gravidade da classe. A figura 2.1 permite uma melhor visualização do método.

A recta que passa pelos pontos 1 e 2 fornece uma indicação da direcção que o novo pólo de agregação deve tomar. De facto, os indivíduos foram agregados à classe de acordo com a proximidade, não com o ponto 2, mas sim com o ponto 1, o que sugere que o movimento do centro de gravidade é retardado, logo o novo centro deve recuperar este atraso "ultrapassando" o centro de gravidade. Jancey considera que este procedimento acelera a convergência do método.

Pode acontecer que o novo centro da classe pertença a outra classe, pelo que neste caso poderá ser importante a implementação de um critério de fusão de classes, como o que será visto na Secção 2.2.3.

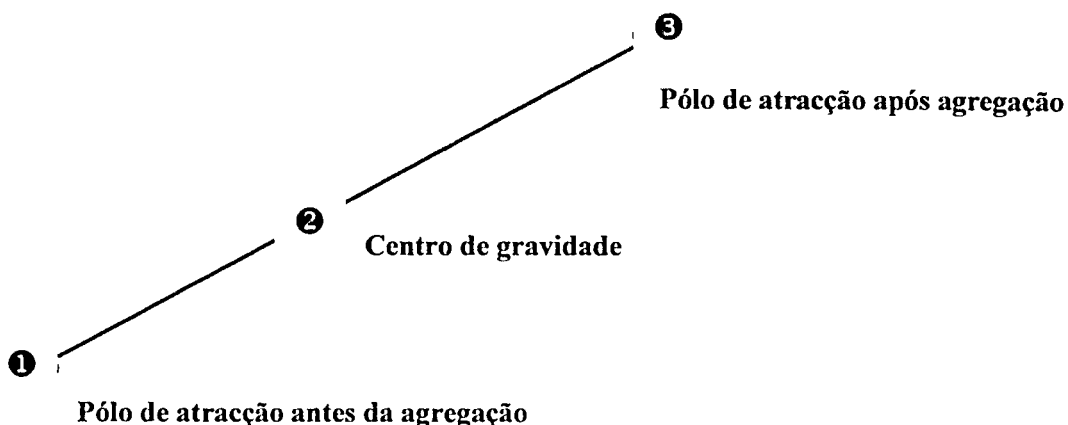


Figura 2.1: Variante de Jancey

2.2.2 Método das k -Médias de MacQueen

MacQueen [28] usa o termo " k -médias" para designar o processo de afectação de cada indivíduo à classe com o centro de gravidade mais próximo. Tal como no método de Forgy, no método de MacQueen, a afectação faz-se com base na distância ao centro mais próximo, no entanto, o algoritmo de MacQueen difere do anterior pelo facto de o centro da classe ser actualizado em cada afectação.

A aplicação do método das k -médias a um conjunto de n elementos, na sua versão mais simples, segue os seguintes passos:

1. Selecção dos k centros iniciais, elementos de E ;
2. Para cada um dos restantes $n - k$ indivíduos:
 - 2.1. Afectação do indivíduo à classe de centro mais próximo.
 - 2.2. Actualização do centro de gravidade da classe.
3. Após a classificação de todos os indivíduos, itera-se o método, reclassificando-os um a um, ajustando em cada passo o centro de gravidade da classe-origem e o da classe-destino.

Usando os k primeiros elementos do conjunto como centros e fazendo, por último, uma única reafectação, este método distingue-se por ser o mais económico de entre os métodos baseados no método de MacQueen. De facto, no total o método requer o cálculo de $k(2n - k)$ distâncias, $(k - 1)(2n - k)$ comparações entre as mesmas e $n - k$ actualizações de centros de gravidade.

cálculo de $k(2n - k)$ distâncias, $(k - 1)(2n - k)$ comparações entre as mesmas e $n - k$ actualizações de centros de gravidade.

2.2.2.1. Propriedades de convergência

Neste parágrafo prova-se que os métodos de Forgy e k -Médias de MacQueen são convergentes. Estabelece-se como critério de convergência a minimização da inércia intra-classes, isto é, uma nova partição é considerada se o total da inércia intra-classes é inferior ao total de inércia intra-classes da partição anterior. Dois pontos essenciais devem ser focados:

- De acordo com o Teorema de Huygens, dada uma classe, a soma dos quadrados dos desvios de cada elemento em relação a um ponto é mínima quando o ponto de referência considerado é o vector centro de gravidade de cada classe, isto é, os momentos centrados são inferiores aos momentos não centrados.
- O número de maneiras de formar k classes não vazias de um conjunto com n elementos é dado pelo número de Stirling de 2ª espécie, expresso em (2.0.1), que é finito $\forall n \in \mathbb{N}$. Assim, qualquer método que gere partições de um conjunto finito é finitamente convergente porque existe um número finito de partições.

Considere-se o valor da inércia intra-classes da partição formada numa determinada etapa do algoritmo,

$$V = \sum_{j=1}^k \sum_{I_i \in A_j} d^2(I_i, g_j),$$

sendo g_j o centro de gravidade da classe A_j e d^2 o quadrado da distância euclideana. Na etapa seguinte, após a afectação dos indivíduos a uma nova classe de cujo centro se encontre mais próximo, considere-se a inércia intra-classes

$$V' = \sum_{j=1}^k \sum_{I_i \in A'_j} d^2(I_i, g_j),$$

sendo A'_j , a nova classe constituída em torno do centro de gravidade da classe A_j , g_j , obtida na etapa anterior. Pretende-se mostrar que $V' < V$.

Seja A'_j uma classe qualquer e $I_\ell \in A'_j$. Se durante o processo de afectação:

- I_ℓ não mudou de classe então, a sua contribuição para V e V' é igual;

- I_l mudou da classe A_h , de centro de gravidade g_h , $h \neq j$, então pelo critério de agregação, $d(I_l, g_j) < d(I_l, g_h)$, logo a contribuição para V' é menor do que a contribuição para V . Aplicando um raciocínio análogo a todas as classes, obtém-se que $V' < V$.

Dado que g_j não é o centro de gravidade da classe A'_j , tem-se que V' não é a inércia intra-classes da nova partição, mas sim,

$$V^* = \sum_{j=1}^k \sum_{I_i \in A'_j} d^2(I_i, g'_j),$$

sendo g'_j o centro de gravidade da classe A'_j . Como já foi referido, o quadrado dos desvios de cada elemento em relação a um ponto é mínima quando o ponto de referência considerado é o vector centro de gravidade de cada classe, portanto $V^* < V' < V$.

Em conclusão, a inércia intra-classes da partição obtida na etapa seguinte diminui, e dado que a sucessão de valores da inércia intra-classes é monótona e limitada, tem-se a convergência do método de Forgy. Como corolário imediato tem-se a convergência do método de MacQueen.

A convergência da partição nos dois métodos foi demonstrada no caso em que se estabelece como critério de convergência a minimização da inércia intra-classes e é usada a distância euclideana. No caso de se ter a distância *city block*, o ponto que minimiza a quantidade

$$\sum_{j=1}^k \sum_{I_i \in A_j} |I_i - c_j|$$

obtém-se quando c_j é a mediana da classe A_j .

2.2.2.2. Método das k -Médias: Forma das Classes

Algumas propriedades do algoritmo, incluindo a convexidade das classes, que de seguida se demonstra, podem ser vistas em Van Ness [41].

Seja P a partição final de um conjunto E de n elementos, descritos por p variáveis, obtida pelo método das k -médias de MacQueen. Considerem-se A_1 e A_2 , duas classes

distintas de P , de centros de gravidade g_1 e g_2 , respectivamente. Se $I_i \in A_1$ então,

$$\begin{aligned} d^2(I_i, g_1) < d^2(I_i, g_2) &\Leftrightarrow \sum_{j=1}^p (x_{ij} - g_{1j})^2 < \sum_{j=1}^p (x_{ij} - g_{2j})^2 \\ &\Leftrightarrow \sum_{j=1}^p \frac{g_{1j}^2 - g_{2j}^2}{2} < \sum_{j=1}^p x_{ij}(g_{1j} - g_{2j}) \end{aligned}$$

sendo, g_{1j} a média da j -ésima variável na classe A_1 , de cardinal n_1 , e g_{2j} a média da j -ésima variável na classe A_2 , de cardinal n_2 . Logo,

$$\sum_{j=1}^p x_{ij}(g_{1j} - g_{2j}) > c \quad (2.2.1)$$

sendo $c = \sum_{j=1}^p \frac{g_{1j}^2 - g_{2j}^2}{2}$. Analogamente, se $I_i \in A_2$, tem-se

$$\sum_{j=1}^p x_{ij}(g_{1j} - g_{2j}) < c. \quad (2.2.2)$$

Em Classificação, uma classe diz-se convexa quando um elemento pertence a uma classe se e só se é uma média ponderada de elementos da mesma classe.

Seja $P = \{A_1, \dots, A_k\}$ uma partição de E . Considere-se um elemento I_ℓ da classe A_2 , que é simultaneamente uma média ponderada de elementos da classe A_1 .

Sendo I_ℓ uma média ponderada de elementos de A_1 , tem-se que ¹

$$I_\ell = \sum_{i=1}^{n_1} w(i) I_i \quad (2.2.3)$$

onde cada componente do vector que representa I_ℓ em $\mathbb{R}^p$, é dada por

$$x_{\ell j} = \sum_{i=1}^{n_1} w(i) x_{ij}, \quad j = 1, \dots, p,$$

e $\sum_{i=1}^{n_1} w(i) = 1$, onde $w(i)$ é o peso de cada indivíduo na média ponderada. Logo,

$$\sum_{j=1}^p x_{\ell j} = \sum_{j=1}^p \sum_{i=1}^{n_1} w(i) x_{ij}.$$

¹Para simplificação de escrita admite-se que os elementos da classe A_1 são os primeiros n_1 indivíduos do conjunto a classificar.

Multiplicando ambos os membros de (2.2.2) por $(g_{1j} - g_{2j})$ e somando em j tem-se

$$\begin{aligned} \sum_{j=1}^p x_{tj}(g_{1j} - g_{2j}) &= \sum_{j=1}^p \sum_{i=1}^{n_1} w(i)x_{ij}(g_{1j} - g_{2j}) \\ &= \sum_{i=1}^{n_1} w(i) \underbrace{\sum_{j=1}^p x_{ij}(g_{1j} - g_{2j})}_{>c}. \end{aligned}$$

Por (2.2.1) obtém-se

$$\begin{aligned} \sum_{j=1}^p x_{tj}(g_{1j} - g_{2j}) &> \sum_{i=1}^{n_1} w(i)c \\ \sum_{j=1}^p x_{tj}(g_{1j} - g_{2j}) &> c \end{aligned}$$

contradizendo (2.2.2).

Por redução ao absurdo provou-se que as classes obtidas pelo método de MacQueen são convexas.

Em problemas unidimensionais, a convexidade de classes significa que estas correspondem a intervalos. Por exemplo, uma partição em duas classes é da forma $\{I_i : x_i < c\}$ e $\{I_i : x_i > c\}$. Para um quadro de dados com duas variáveis, uma partição em duas classes implica que as classes são separadas por rectas; em problemas de ordem superior, as classes são separadas por hiperplanos.

2.2.3 Método k -Médias de MacQueen: Fusão e Refinamento

MacQueen [28] propõe uma variante ao método das k -médias, que durante a agregação dos indivíduos permite a adaptação do número de classes. Consta fundamentalmente dos seguintes passos:

1. Definem-se os parâmetros k (número de classes), β (parâmetro de fusão) e α (parâmetro de refinamento).
2. Seleccionam-se os k primeiros elementos da amostra para centros iniciais. Calculam-se todas as distâncias entre pares de centros iniciais. Se a menor destas distâncias for inferior ao parâmetro de fusão β , agrupam-se os centros correspondentes à menor distância e calcula-se o centro de gravidade da classe formada

por estes dois elementos. Prossegue-se calculando a distância entre o centro desta classe e os restantes $k - 2$ centros. A agregação entre classes é finalizada quando todos os centros se encontrarem a uma distância superior ou igual a β .

3. Agregam-se os restantes $n - k$ indivíduos à classe cujo centro se encontra mais próximo. Em cada agregação, actualiza-se o centro da classe-origem e da classe-destino e calcula-se a distância entre o centro da nova classe e as restantes classes. Se a distância entre algum dos $n - k$ indivíduos ao centro mais próximo for superior ao parâmetro de refinamento α , forma-se uma nova classe com este único elemento, aumentando assim o número de classes.
4. Comparam-se os centros de cada classe: se o centro de gravidade se encontrar a uma distância inferior a β de algum centro de outra classe, agregam-se as duas classes.

Itera-se o algoritmo até não haver alteração na constituição das classes.

Uma das vantagens da variante aqui apresentada incide no facto de a fusão de classes com centros de gravidade próximos impedir a divisão de classes razoavelmente homogêneas. Além disso, o processo de refinamento permite que, de forma natural, os elementos atípicos (*outliers*) formem novas classes em vez de se forçar a sua entrada em classes já existentes. Ultrapassa-se, assim, a eventual arbitrariedade na escolha de k .

2.2.4 Método k -Médias de MacQueen: Variante de Wishart

David Wishart (cf. Anderberg [1]) introduz no método de MacQueen sistemáticas reduções no número de classes. Este procedimento tem por base o cálculo das distâncias entre pares de classes e, posteriormente, a eventual fusão das duas classes mais próximas.

O algoritmo compreende os seguintes passos:

1. Fixam-se os valores dos parâmetros de controlo: Δ (distância máxima admissível dos indivíduos aos centros), *mini* (número mínimo de elementos de cada classe), *maxit* (número máximo de iterações permitido) e *nc* (número mínimo de classes pretendido).

2. Define-se uma partição inicial com k classes ($k \geq nc$) e calcula-se o centro de gravidade de cada classe.
3. Calcula-se a distância de cada um dos indivíduos aos k centros de gravidade, $d(I_i, g_j)$, $i = 1, \dots, n$, $j = 1, \dots, k$.

- i. Se a menor distância é superior a Δ , isto é,

$$\min d(I_i, g_j) > \Delta,$$

então, agrega-se o indivíduo a um *grupo resíduo* de indivíduos não classificados até ao momento, e actualiza-se o centro de gravidade da classe-origem do indivíduo.

- ii. Se o centro de gravidade mais próximo, $g_{j'}$, não é o centro de gravidade da classe-origem do indivíduo e a distância a $g_{j'}$ é inferior ou igual a Δ , então coloca-se o indivíduo na classe de centro $g_{j'}$. Imediatamente actualiza-se o centro de gravidade da classe-origem e da classe-destino.
- iii. Se o indivíduo se encontra no *grupo resíduo* e a sua distância ao centro de gravidade mais próximo for inferior ou igual a Δ , afecta-se o indivíduo à respectiva classe e actualiza-se o centro de gravidade da classe-destino.
4. Após a afectação de todos os indivíduos do conjunto tal como sugere o passo 3, agregam-se ao *grupo resíduo* as classes com menos de *mini* indivíduos, o que eventualmente reduzirá o número de classes.
5. Repetem-se os passos 3 e 4 até convergência do algoritmo, ou o número máximo de iterações permitido *maxit* ser atingido.
6. Calcula-se a semelhança entre todos os pares de classes e agregam-se as duas classes mais próximas. Repetem-se os passos 2 a 5. O algoritmo é finalizado quando o número de classes obtido é nc .

Duas considerações sobre o algoritmo proposto por Wishart parecem ser importantes para mostrar a sua eficácia. A primeira é a de que *outliers* severos podem ser facilmente identificados se estes permanecerem no *grupo resíduo*. A outra vantagem do algoritmo prende-se com o uso do parâmetro de controlo *maxit* o qual se torna particularmente útil na presença de *indivíduos cíclicos*, isto é, indivíduos que entram e saem

do grupo *resíduo*, impedindo a convergência do método.

2.2.5 Método k -Medóides

Outro método de Classificação Não Hierárquica, de tipo centróide, foi proposto por Vinod (cf. Kaufman e Rousseeuw [24]): método k -medóides. Este método engloba todos os passos integrantes do algoritmo das k -médias de MacQueen, diferindo apenas na definição do representante de cada classe: o centro da classe deixa de ser o seu centro de gravidade para dar lugar a um elemento do próprio grupo. No método k -medóides, o centro de uma classe é o elemento mais central relativamente aos restantes elementos da classe, o que permite a minimização média das dissemelhanças de todos os elementos relativamente ao elemento representativo da classe na qual se inserem. Atribui-se o termo *medóide* ao elemento da classe cuja dissemelhança média relativamente aos restantes elementos desta, é mínima.

Sejam c_j , com $j = 1, \dots, k$, k elementos de E , representativos das k classes. O modelo subjacente ao método k -medóides assenta em minimizar globalmente a dissemelhança entre cada elemento e o elemento representativo da respectiva classe, isto é, minimiza-se ²

$$\sum_{j=1}^k \sum_{\ell=k+1}^n \gamma(c_j, I_\ell) \text{Ind}_{c_j, I_\ell} \quad (2.2.4)$$

sendo γ uma função de dissemelhança e Ind_{c_j, I_ℓ} a função indicatriz dos elementos representativos, isto é, a função toma o valor 1 se I_ℓ pertence à classe cujo elemento representativo é c_j , caso contrário, assume o valor 0.

Tendo em conta a minimização da função objectivo (2.2.4), um algoritmo que permite a escolha dos primeiros k pontos representativos é proposto por Kaufman e Rousseeuw [24]. Este contempla duas etapas: selecção dos k pontos representativos do conjunto E e validação da escolha realizada.

Numa primeira fase tem-se:

²Para simplificação de escrita considera-se que os elementos representativos das k classes são os k primeiros indivíduos do conjunto a classificar.

1. O primeiro centro c_1 , é o ponto cuja soma das dissemelhanças a todos os restantes indivíduos é mínima, ou seja,

$$c_1 \equiv I_t \in E : \sum_{i=1}^n \gamma(I_t, I_i) = \min \left\{ \sum_{i=1}^n \gamma(I_{t'}, I_i) : I_{t'} \in E \right\}.$$

Desta forma, c_1 é o elemento mais central do conjunto de dados.

2. Os elementos representativos seguintes obtêm-se de acordo com o seguinte procedimento:

- i. Considera-se um elemento I_i ainda não seleccionado para elemento representativo. Escolhe-se outro elemento I_t na mesma condição de I_i .
- ii. Representando por D_{I_t} a dissemelhança entre I_t e o elemento representativo mais próximo já seleccionado, calcula-se o contributo de I_t na escolha de I_i para elemento representativo,

$$C_{it} = \max\{D_{I_t} - \gamma(I_i, I_t), 0\}.$$

Se $C_{it} > 0$, I_t contribui para a decisão de seleccionar I_i para elemento representativo.

- iii. Calcula-se o total da contribuição no caso de se escolher I_i para j -ésimo centro,

$$\sum_{I_t \in E \setminus \{c_1, \dots, c_{j-1}, I_i\}} C_{it}. \quad (2.2.5)$$

Selecciona-se para j -ésimo elemento representativo, com $j = 2, \dots, k$, o elemento I_i que maximiza (2.2.5), ou seja,

$$\begin{aligned} c_j &\equiv I_i \in E : \sum_{I_t \in E \setminus \{c_1, \dots, c_{j-1}, I_i\}} C_{it} \\ &= \max \left\{ \sum_{I_t \in E \setminus \{c_1, \dots, c_{j-1}, I_{t'}\}} C_{it'} : I_{t'} \in E \setminus \{c_1, \dots, c_{j-1}\} \right\}. \end{aligned}$$

Numa segunda fase, valida-se a escolha dos k representantes das classes, procedendo-se a uma substituição de algum ponto no caso do procedimento que se segue o justificar. Considerem-se todos os pares de elementos (c, I_i) , $c \in \{c_1, \dots, c_k\}$ e $I_i \in E \setminus \{c_1, \dots, c_k\}$. Para verificar se é razoável a troca entre os papéis destes dois elementos, isto é, I_i passar a ser um ponto representativo em desfavor de c , são levados a cabo os seguintes cálculos:

1. Considere-se um elemento não representativo I_ℓ . Represente-se por C_{tic} , o contributo de I_ℓ para a minimização da quantidade em (2.2.4), no caso de se verificar a troca entre I_i e c . Tem-se então que:

- a) se I_ℓ é mais distante de I_i e de c do que qualquer outro elemento representativo c' , isto é, $\gamma(I_\ell, I_i) > \gamma(I_\ell, c')$ e $\gamma(I_\ell, c) > \gamma(I_\ell, c')$, $c' \in \{c_1, \dots, c_k\} \setminus \{c\}$, então obviamente neste caso não se contribui para a minimização da função objectivo definida em (2.2.4), logo, $C_{tic} = 0$.
- b) se c é o elemento representativo mais próximo de I_ℓ , isto é, $D_{I_\ell} = \gamma(I_\ell, c)$, então duas situações podem ocorrer:
 - b1) I_ℓ encontra-se mais próximo de I_i do que o segundo centro mais próximo de I_ℓ , c' , isto é, $\gamma(I_\ell, I_i) < \gamma(I_\ell, c')$. Neste caso, a contribuição de I_ℓ para a troca entre I_i e c é dada por

$$C_{tic} = \gamma(I_\ell, I_i) - \gamma(I_\ell, c).$$

Note-se que neste caso, se I_ℓ se encontra mais próximo de c do que de I_i , o elemento I_ℓ não contribui para a troca entre I_i e c . Apenas quando $C_{tic} < 0$, é favorável a troca entre I_i e c .

- b2) I_ℓ encontra-se a uma distância de I_i igual ou superior à de c' , segundo elemento representativo mais próximo de I_ℓ , isto é, $\gamma(I_\ell, I_i) \geq \gamma(I_\ell, c')$. Neste caso,

$$C_{tic} = \gamma(I_\ell, I_i) - \gamma(I_\ell, c).$$

O valor de C_{tic} é sempre positivo, o que indica que a troca entre I_i e c não se deve realizar.

- c) se I_ℓ é mais distante de c do que de algum dos elementos representativos, mas mais próximo de I_i do que qualquer centro, então

$$C_{tic} = \gamma(I_\ell, I_i) - \gamma(I_\ell, c'),$$

sendo c' o elemento representativo mais próximo de I_ℓ .

2. Calcula-se o total das contribuições C_{tic}

$$T_{ci} = \sum_{I_\ell \in E \setminus \{c_1, \dots, c_k, I_i\}} C_{tic}.$$

3. Escolhe-se o par de elementos (c, I_i) que minimiza T_{ci} .
4. A troca entre c e I_i é realizada se T_{ci} é um valor negativo, retomando-se de seguida, o passo 1.

Note-se que a escolha dos pontos representativos pela ordem de recolha é ultrapassada, uma vez que o algoritmo apresentado permite estudar todas as potenciais trocas a efectuar.

2.3 Método das Nuvens Dinâmicas

O método das Nuvens Dinâmicas (Diday [11]) tal como os métodos apresentados neste capítulo, é um método de tipo centróide, no entanto distingue-se dos anteriores por representar simbolicamente um conjunto de indivíduos por um "núcleo". Esta representação simbólica pode ser um ponto, elemento ou não do conjunto a classificar, um conjunto de pontos, um centro de gravidade, uma recta, ou uma função. Diferentes abordagens foram desenvolvidas considerando diferentes espaços de representação: centróides (Diday e Simon [12]), eixos factoriais (Bock [8], Ok-Sakun [34]), leis de probabilidade (Schoeder [39]), etc.

Outra característica inovadora do método das Nuvens Dinâmicas é a de definir um critério a otimizar que permita medir a adequação entre uma classificação e a representação dessa classificação. Assim, o problema reside na procura simultânea de uma classificação e respectiva representação que optimizem localmente o critério. Duas funções têm um papel fundamental neste método:

- a **função de afectação**, que permite obter as classes a partir de um conjunto de núcleos;
- a **função de representação**, que inversamente à função de afectação, permite a partir de uma partição do conjunto a classificar obter um conjunto de núcleos.

O algoritmo das Nuvens Dinâmicas considera um critério, W , que mede a adequação entre as classes e os respectivos núcleos. Sendo $\mathbb{L}_k = \mathbb{L}^k$ o conjunto dos k -uplos de núcleos admissíveis, e $\mathbb{P}_k$ o conjunto das partições em k classes de E , W é uma aplicação

$$W : \mathbb{L}_k \times \mathbb{P}_k \longrightarrow \mathbb{R}^+$$

definida por

$$W(N, P) = \sum_{j=1}^k D(N_j, A_j)$$

onde D é uma medida de dissemelhança. Um baixo valor de D , significa que o núcleo N_j se adequa à classe A_j .

Sejam A e B classes do conjunto E a classificar e γ uma medida do tipo dissemelhança. Admita-se que

$$D(A, B) = \sum_{a \in A} \sum_{b \in B} \gamma(a, b),$$

isto é, $D(A, B)$ é o somatório das dissemelhanças entre pares de elementos de $A \times B$.

Seguem-se as definições de função de afectação e de função de representação.

- **Função de Afectação**

A função de afectação f é uma aplicação de $\mathbb{L}_k$ em $\mathbb{P}_k$,

$$f : \mathbb{L}_k \longrightarrow \mathbb{P}_k.$$

Dado o vector de núcleos $(N_1, \dots, N_k)$, a partição $P = \{A_1, \dots, A_k\}$ é dada por

$$A_j = \{I_\ell \in E : D(N_j, I_\ell) \leq D(N_{j'}, I_\ell), 1 \leq j' \leq k\}, j = 1, \dots, k.$$

Se $D(N_j, I_\ell) = D(N_{j'}, I_\ell)$ afecta-se I_ℓ à classe de mais baixo índice.

- **Função de Representação**

A função de representação g é uma aplicação de $\mathbb{P}_k$ em $\mathbb{L}_k$,

$$g : \mathbb{P}_k \longrightarrow \mathbb{L}_k.$$

Dada uma partição $P = \{A_1, \dots, A_k\}$ de E , o vector de núcleos $(N_1, \dots, N_k)$ obtêm-se do seguinte modo:

$$N_j = L \in \mathbb{L} : D(L, A_j) \text{ minimiza } D(\bullet, A_j).$$

Admite-se que o elemento $L \in \mathbb{L}$ que minimiza $D(\bullet, A_j)$ existe e é único.

Aplicando alternadamente as funções f e g a partir de um conjunto de k núcleos iniciais, pretende-se obter um par $(N^*, P^*) \in \mathbb{L}_k \times \mathbb{P}_k$ que minimize o critério

$$W : \mathbb{L}_k \times \mathbb{P}_k \longrightarrow \mathbb{R}^+,$$

$$(N^*, P^*) : W(N^*, P^*) = \min\{W(N, P) : (N, P) \in \mathbb{L}_k \times \mathbb{P}_k\}.$$

Resumidamente, o método das Nuvens Dinâmicas consta essencialmente dos seguintes passos:

1. Escolha de um espaço de representação $\mathbb{L}_k$.
2. Definição do critério $W : \mathbb{L}_k \times \mathbb{P}_k \longrightarrow \mathbb{R}^+$ que avalia o ajustamento entre um vector de núcleos e uma partição em k classes.
3. Definição do problema de optimização a fim de minimizar o critério W .
4. Construção de um algoritmo que consiste basicamente na aplicação alternada da função de representação e da função de afectação.

O algoritmo das Nuvens Dinâmicas pode ser formalizado recorrendo às sucessões

$$v_m = (N^{(m)}, P^{(m)}) \in \mathbb{L}_k \times \mathbb{P}_k,$$

$$N^m = g(P^{(m-1)}) \text{ e } P^{(m)} = f(g(P^{(m-1)}))$$

e

$$u_m = W(v_m).$$

O algoritmo calcula os diferentes termos da sucessão v_m , até convergência. Em Diday [11] prova-se a convergência das sucessões u_m e v_m .

2.3.1 Variante dos Centros de Gravidade

Esta variante corresponde ao método de Forgy: cada classe é representada pelo seu centro de gravidade.

• Espaço de Representação das Classes

Considere-se que os indivíduos são representados por vectores de $\mathbb{R}^p$, munido da distância euclideana d . O espaço de representação $\mathbb{L}$ de uma classe é também o espaço $\mathbb{R}^p$. Define-se

$$D(\{I_\ell\}, A) = \sum_{I_i \in A} p_i d^2(I_\ell, I_i) = \text{Iner}(I_\ell, A), \forall A \in P(E), \forall I_\ell \in \mathbb{R}^p$$

onde p_i é o peso de I_i e $\text{Iner}(I_\ell, A)$ é a inércia de A em relação a I_ℓ .

- **Função de representação**

O problema de otimização consiste em minimizar $D(\{x\}, A)$, $x \in \mathbb{R}^p$, que de acordo com o Teorema de Huygens, admite uma única solução que é o centro de gravidade da classe A .

A função de representação $g : \mathbb{P}_k \rightarrow \mathbb{L}_k$ é definida por $g(\{A_1, \dots, A_k\}) = (N_1, \dots, N_k)$, sendo N_j o centro de gravidade de A_j , com $j = 1, \dots, k$.

- **Função de Afecção**

A função de afecção $f : \mathbb{L}_k \rightarrow \mathbb{P}_k$ é definida por $f(N_1, \dots, N_k) = \{A_1, \dots, A_k\}$, sendo

$$A_j = \{I_\ell \in E : D(N_j, \{I_\ell\}) \leq D(N_{j'}, \{I_\ell\}), 1 \leq j' \leq k\}$$

em caso de igualdade I_ℓ será afectado à classe de menor índice.

- **O Problema de Otimização**

Pretende-se encontrar o melhor par $(N, P) \in \mathbb{L}_k \times \mathbb{P}_k$ que minimize o critério de adequação entre a partição $P = \{A_1, \dots, A_k\}$ e o vector de representantes $N = (N_1, \dots, N_k)$ definido por

$$W(N, P) = \sum_{j=1}^k D(N_j, A_j) = \sum_{j=1}^k \sum_{I_i \in A_j} d^2(N_j, I_i).$$

Como o representante de cada classe A_j é o seu centro de gravidade g_j , o critério W é dado por

$$W(N, P) = \sum_{j=1}^k D(N_j, A_j) = \sum_{j=1}^k \sum_{I_i \in A_j} d^2(I_i, g_j) = \sum_{j=1}^k Iner_j$$

onde $Iner_j$ é a inércia da classe A_j em relação ao seu centro de gravidade g_j .

Desta forma, o critério $W(N, P)$ é a inércia intra-classe da partição P .

O método das Nuvens Dinâmicas maximiza a soma dos quadrados dos desvios entre as médias das variáveis nas classes e as médias das variáveis globais. Minimizando a inércia intra-classe, o método das Nuvens Dinâmicas fornece classes esféricas, de volume semelhante e de baixa inércia.

Capítulo 3

Modelos Probabilísticos em Classificação Não Hierárquica

A procura de métodos de classificação que satisfaçam os diversos problemas oriundos de diferentes áreas, tem dado origem a um amplo conjunto de procedimentos classificatórios. Em particular, os métodos aqui propostos inserem-se na metodologia VL (Validade da Ligação) que se caracterizam por recorrerem à medida de probabilidade para definir as funções de comparação.

Como já foi referido no Capítulo 1, os primeiros métodos desenvolvidos, no quadro da Classificação Hierárquica, segundo esta metodologia são os métodos AVL, cuja função de comparação entre classes é dada pelo valor da função de distribuição do máximo dos valores das semelhanças entre pares de elementos, e o método AVM, no qual a medida de comparação é o valor da função de distribuição da média aritmética dos valores da função de comparação entre os pares de elementos das classes. O método AVL proposto por Lerman [26], foi amplamente estudado por Bacelar-Nicolau [3]. Posteriormente, o método AVM foi proposto e estudado por Nicolau [29]. Mais recentemente, outros métodos VL baseados em estatísticas de tendência central foram propostos por Sousa [40]: o recurso à média geométrica, ao centro e à mediana, deram origem, em Classificação Hierárquica, aos métodos **AVMG** (Agregação pela Validade da Média Geométrica), **AVC** (Agregação pela Validade do Centro) e **AVMED** (Agregação pela Validade da MEDiana). A adaptação dos métodos VL de Classificação Hierárquica à Classificação Não Hierárquica surge naturalmente: de facto, em cada etapa de afectação de um elemento a uma classe procede-se à reunião de duas classes, uma das quais

tem cardinal igual a um. Em Classificação Não Hierárquica, foi desenvolvido com a abordagem VL o método NHMEAN (Non Hierarchical MEAN), cuja função de afectação é dada pela função de distribuição da estatística média aritmética. O método NHMEAN, explorado por Nicolau e Brito ([9], [33]), forneceu resultados muito satisfatórios.

O estudo de métodos classificatórios que recorrem a estatísticas de tendência central, em Classificação Não Hierárquica, está longe de se encontrar esgotado, pelo que com este trabalho se pretende dar algum contributo a esta problemática. Duas abordagens são feitas neste capítulo. A primeira, aqui designada por abordagem probabilística, usa na afectação dos elementos às classes os valores da função de distribuição das estatísticas amostrais dos valores do coeficiente de semelhança VL. Neste contexto, são apresentados quatro métodos de Classificação Não Hierárquica que têm em comum o facto de usarem a função de distribuição de uma estatística de tendência central dos valores do coeficiente de semelhança VL. Numa segunda abordagem, os métodos de Classificação Não Hierárquica, designados por métodos directos, resultam da aplicação directa de estatísticas amostrais de tendência central aos valores do coeficiente de semelhança VL.

3.1 Métodos VL: Abordagem probabilística

Considere-se o conjunto E a classificar e seja $\gamma : E \times E \rightarrow [0, 1]$ a aplicação escolhida como medida de comparação entre elementos de E do tipo semelhança. Considerando, como hipótese de referência, a independência entre os elementos de E , admite-se que o conjunto

$$\{\gamma(I_i, I_\ell) : I_i, I_\ell \in E, i \neq \ell\} \quad (3.1.1)$$

constitui uma amostra de $\frac{n(n-1)}{2}$ variáveis aleatórias independentes e identicamente distribuídas com lei uniforme em $[0, 1]$. Assim, toda a informação de que se dispõe para dar início ao processo de classificação está disponível no conjunto em (3.1.1).

Os critérios de agregação baseados em modelos probabilísticos assentam no uso de diferentes estatísticas sobre o conjunto de valores em (3.1.1), decorrendo que para medir a proximidade entre classes serão utilizados os valores das funções de distribuição

dessas estatísticas, sempre na hipótese de referência. Os novos métodos aqui propostos assentam no uso das estatísticas amostrais de tendência central, mediana, média geométrica e centro. As funções de distribuição das estatísticas subjacentes aos métodos que são expostos nos parágrafos seguintes foram já deduzidas no quadro da Classificação Hierárquica por Sousa [40]. Como já foi referido, a adaptação à Classificação Não Hierárquica é particularmente facilitada: em cada etapa de afectação procede-se à reunião de duas classes, uma das quais tem cardinal igual a um. Assim, aos métodos de Classificação Hierárquica **AVM** (Agregação pela Validade da Média), **AVMED** (Agregação pela Validade da MEDiana), **AVMG** (Agregação pela Validade da Média Geométrica) e **AVC** (Agregação pela Validade do Centro), correspondem, respectivamente, em Classificação Não Hierárquica, os métodos **NHMEAN** (Non Hierarchical MEAN), **NHVMED** (Non Hierarchical Validity of the MEDian), **NHVGM** (Non Hierarchical Validity of the Geometric Mean) e **NHVC** (Non Hierarchical Validity of the Center).

A exposição de cada um dos métodos de Classificação Não Hierárquica engloba o respectivo critério de agregação em Classificação Hierárquica, bem como alguns pontos essenciais na dedução da função de agregação. Ao longo do trabalho é usada a notação γ para designar um índice de comparação do tipo semelhança, caso contrário, será previamente esclarecido que γ é do tipo dissemelhança. Para o que se segue denota-se por x um elemento de E , conjunto a classificar.

3.1.1 Método NHMEAN

Em Classificação Hierárquica, o método AVM (Nicolau [29]) associado ao uso da estatística U_M , média aritmética amostral, usa como medida de comparação entre duas classes A e B , a função

$$\Gamma_{AVM}(A, B) =$$

$$\begin{cases} 0 & \text{se } z < 0 \\ \frac{1}{n_A n_B!} \sum_{j=0}^{n_A n_B z} (-1)^j \binom{n_A n_B}{j} (n_A n_B z - j)^{n_A n_B} & \text{se } 0 \leq z \leq 0.5, n_A n_B \leq 11 \\ 1 - \frac{1}{n_A n_B!} \sum_{j=0}^{n_A n_B (1-z)} (-1)^j \binom{n_A n_B}{j} (n_A n_B (1-z) - j)^{n_A n_B} & \text{se } 0.5 < z \leq 1, n_A n_B \leq 11 \\ \Phi(\sqrt{3n_A n_B}(2z - 1)) & \text{se } 0 < z \leq 1, n_A n_B \geq 12 \\ 1 & \text{se } z > 1. \end{cases}$$

$$\text{sendo } z = u_M(A, B) = \frac{\sum_{(a,b) \in A \times B} \gamma(a, b)}{n_A n_B}, \text{ concretização da estatística } U_M.$$

Em Classificação Não Hierárquica, designa-se por NHMEAN o método que tem por base a estatística média aritmética amostral. Sendo uma das classes constituída por um único elemento, o critério de afectação de um elemento x a uma classe A fica definido pela expressão

$$\Gamma_{NHMEAN}(\{x\}, A) =$$

$$= \begin{cases} 0 & \text{se } z < 0 \\ \frac{1}{n_A!} \sum_{i=0}^{n_A z} (-1)^i \binom{n_A}{i} (n_A z - i)^{n_A} & \text{se } 0 \leq z \leq 0.5, n_A \leq 11 \\ 1 - \frac{1}{n_A!} \sum_{i=0}^{n_A (1-z)} (-1)^i \binom{n_A}{i} (n_A (1-z) - i)^{n_A} & \text{se } 0.5 < z \leq 1, n_A \leq 11 \\ \Phi(\sqrt{3n_A}(2z - 1)) & \text{se } 0 < z \leq 1, n_A \geq 12 \\ 1 & \text{se } z > 1 \end{cases} \quad (3.1.2)$$

$$\text{para } z = u_M(\{x\}, A) = \frac{\sum_{a \in A} \gamma(x, a)}{n_A}.$$

3.1.2 Método NHVGM

O uso da estatística média geométrica no critério de agregação entre duas classes deu lugar, em Classificação Hierárquica, ao método AVMG.

A estatística U_{MG} , média geométrica dos valores de comparação dos pares de elementos de duas classes A e B , de cardinais n_A e n_B , respectivamente, é dada por

$$U_{MG} = \sqrt[n_A n_B]{\prod_{(a,b) \in A \times B} \gamma(a,b)} \quad (3.1.3)$$

Aplicando logaritmos neperianos a (3.1.3) e considerando o seu simétrico obtém-se:

$$-n_A n_B \ln U_{MG} = \sum_{(a,b) \in A \times B} -\ln \gamma(a,b). \quad (3.1.4)$$

Na hipótese de referência de independência, $\gamma(a,b)$ são variáveis aleatórias independentes e uniformemente distribuídas em $[0, 1]$; tem-se então que $-\ln \gamma(a,b)$ segue uma distribuição exponencial de parâmetro 1, $\text{Exp}(1)$. Portanto, em (3.1.4) está-se perante a soma de $n_A n_B$ variáveis aleatórias independentes cuja distribuição é $\text{Exp}(1)$, que é uma variável aleatória com distribuição gama de parâmetros $n_A n_B$ e 1,

$$-n_A n_B \ln U_{MG} \sim G(n_A n_B, 1). \quad (3.1.5)$$

Tratando-se o primeiro parâmetro $n_A n_B$, de um inteiro, e o segundo parâmetro, da unidade, a distribuição assume a forma de uma lei qui-quadrado, com $n_A n_B$ graus de liberdade, logo,

$$-2n_A n_B \ln U_{MG} \sim \chi^2(2n_A n_B). \quad (3.1.6)$$

Tomando $u_{MG}(A, B)$ como a realização da estatística média geométrica U_{MG} para o par de classes (A, B) , a função de comparação entre A e B que está na base do método AVMG é dada por:

$$\begin{aligned} \Gamma_{VMG}(A, B) &= P(U_{MG} \leq u_{MG}(A, B)) \\ &= P(-2n_A n_B \ln U_{MG} \geq -2n_A n_B \ln u_{MG}(A, B)) \\ &= 1 - F(-2n_A n_B \ln u_{MG}(A, B)) \end{aligned} \quad (3.1.7)$$

sendo F a função de distribuição da $\chi^2(2n_A n_B)$.

No quadro da Classificação Não Hierárquica, surge o método NHVGM no qual se atribui x à classe A que maximiza o valor da função objectivo

$$\Gamma_{NHVGM}(\{x\}, A) = 1 - F(-2n_A \ln u_{MG}(\{x\}, A)), \quad (3.1.8)$$

sendo F a função de distribuição da $\chi^2(2n_A)$ e

$$u_{MG}(\{x\}, A) = \sqrt[n_A]{\prod_{a \in A} \gamma(x, a)}.$$

3.1.3 Método NHVMED

O critério que é aqui referido baseia-se na estatística de ordem mediana dos valores de comparação entre pares de elementos, denotada por U_{Med} . Tem-se então que para duas classes A e B ,

$$U_{Med} = \text{mediana } \{\gamma(a, b) : (a, b) \in A \times B\}. \quad (3.1.9)$$

Em Classificação Hierárquica, designa-se por AVMED o método VL cujo critério de agregação tem por base a mediana amostral. A expressão da função de distribuição da estatística ordinal mediana depende do facto de a dimensão da amostra ser par ou ímpar.

Nas duas sub-seções seguintes segue-se de perto o que foi feito em Sousa [40].

Dimensão Amostral Ímpar

A mediana de um conjunto ordenado de dimensão N ímpar, $\{v_{(1:N)}, v_{(2:N)}, \dots, v_{(N:N)}\}$ é o elemento central do mesmo. Se $N = 2r + 1$, com r inteiro, é a dimensão da amostra, a mediana é a estatística de ordem $r + 1$, $v_{(r+1:N)}$.

O resultado que de seguida se enuncia é usado na dedução da função de distribuição da estatística mediana de uma amostra de dimensão ímpar. Para tal, considerem-se $X_1, X_2, \dots, X_N$, N variáveis aleatórias contínuas, independentes, de distribuição $F(x)$,

$X_{(1:N)}, X_{(2:N)}, \dots, X_{(N:N)}$, as correspondentes variáveis aleatórias ordinais, e represente-se a função densidade de probabilidade e a função de distribuição da estatística de ordem t , por $f_{(t:N)}$ e $F_{(t:N)}$, respectivamente, com $t = 1, \dots, N$.

Proposição 3.1.3.1 Para $t \in \{1, \dots, N\}$, a função densidade de probabilidade da variável aleatória ordinal $X_{(t:N)}$ é dada por:

$$f_{(t:N)}(x) = \frac{N!}{(t-1)!(N-t)!} F(x)^{t-1} [1 - F(x)]^{N-t} f(x) \quad (3.1.10)$$

sendo $f(x)$ e $F(x)$, a função densidade de probabilidade e a função de distribuição da variável aleatória X_t , respectivamente.

Na hipótese de referência, $\{\gamma(a, b) : (a, b) \in A \times B\}$ constitui um conjunto de variáveis aleatórias independentes com distribuição uniforme em $[0, 1]$. Aplicando o resultado expresso em (3.1.10), a função densidade de probabilidade de $X_{(t:N)}$ é dada por:

$$f_{(t:N)}(x) = \begin{cases} \frac{N!}{(t-1)!(N-t)!} x^{t-1} (1-x)^{N-t} & 0 \leq x \leq 1 \\ 0 & \text{c. c.} \end{cases} \quad (3.1.11)$$

Sendo t e N inteiros positivos, está-se perante uma lei beta, com parâmetros t e $N - t + 1$,

$$X_{(t:N)} \sim \text{Beta}(t, N - t + 1).$$

No quadro da Classificação Hierárquica, sendo A e B duas classes de cardinais n_A e n_B , $n_A n_B$ um número ímpar de pares $(a, b) \in A \times B$, então U_{Med} segue uma lei beta de parâmetros $\frac{n_A n_B + 1}{2}$, isto é,

$$U_{Med} \sim \text{Beta}\left(\frac{n_A n_B + 1}{2}, \frac{n_A n_B + 1}{2}\right).$$

Usando (3.1.11), a fórmula do critério AVMED que exprime a semelhança entre as classes A e B , é dada por

$$\Gamma_{VMED}(A, B) =$$

$$= \begin{cases} 0 & \text{se } u_{Med}(A, B) < 0 \\ \frac{n_A n_B!}{\left(\frac{n_A n_B - 1}{2}\right)! \left(\frac{n_A n_B - 1}{2}\right)!} \int_0^{u_{Med}(A, B)} t^{\frac{n_A n_B - 1}{2}} (1-t)^{\frac{n_A n_B - 1}{2}} dt & \text{se } 0 \leq u_{Med}(A, B) \leq 1 \\ 1 & \text{se } u_{Med}(A, B) > 1 \end{cases}$$

$$= \begin{cases} 0 & \text{se } u_{Med}(A, B) < 0 \\ \frac{B_{u_{Med}(A, B)}\left(\frac{n_A n_B + 1}{2}, \frac{n_A n_B + 1}{2}\right)}{B\left(\frac{n_A n_B + 1}{2}, \frac{n_A n_B + 1}{2}\right)} & \text{se } 0 \leq u_{Med}(A, B) \leq 1 \\ 1 & \text{se } u_{Med}(A, B) > 1 \end{cases}$$

considerando as funções beta de parâmetros $\frac{n_A n_B + 1}{2}$ inteiros dadas por

$$B_{u_{Med}(A, B)}\left(\frac{n_A n_B + 1}{2}, \frac{n_A n_B + 1}{2}\right) = \int_0^{u_{Med}(A, B)} t^{\frac{n_A n_B - 1}{2}} (1-t)^{\frac{n_A n_B - 1}{2}} dt$$

e

$$B\left(\frac{n_A n_B + 1}{2}, \frac{n_A n_B + 1}{2}\right) = \int_0^1 t^{\frac{n_A n_B - 1}{2}} (1-t)^{\frac{n_A n_B - 1}{2}} dt = \frac{\left(\frac{n_A n_B - 1}{2}\right)! \left(\frac{n_A n_B - 1}{2}\right)!}{(n_A n_B)!}.$$

Remetendo para a Classificação Não Hierárquica, se n_A é um número ímpar, x é incluído na classe A que torna máximo o valor da expressão

$$\begin{aligned} \Gamma_{NHVMED}(\{x\}, A) &= \frac{B_{u_{Med}(\{x\}, A)}\left(\frac{n_A + 1}{2}, \frac{n_A + 1}{2}\right)}{B\left(\frac{n_A + 1}{2}, \frac{n_A + 1}{2}\right)} \\ &= \frac{n_A!}{\left(\frac{n_A - 1}{2}\right)! \left(\frac{n_A - 1}{2}\right)!} \int_0^{u_{Med}(\{x\}, A)} t^{\frac{n_A - 1}{2}} (1-t)^{\frac{n_A - 1}{2}} dt \end{aligned} \quad (3.1.12)$$

tomando

$$u_{Med}(\{x\}, A) = \text{mediana}\{\gamma(x, a) : a \in A\}.$$

Dimensão Amostral Par

Perante uma amostra ordenada $\{v_{(1:N)}, v_{(2:N)}, \dots, v_{(N:N)}\}$ de dimensão par, $N = 2r$, r inteiro, a estatística mediana é a média aritmética das estatísticas de ordem r e $r+1$, isto é, $U_{Med} = \frac{v_{(r:N)} + v_{(r+1:N)}}{2}$.

Considerando A e B , duas classes de dimensão n_A e n_B , respectivamente, se $n_A n_B$ é um número par, no critério AVMED a semelhança entre as duas classes é traduzida pelo valor da função de distribuição no ponto $u_{Med}(A, B)$ definida através da expressão analítica (Sousa [40])

$$\Gamma_{VMED}(A, B) =$$

$$= \begin{cases} 0 & \text{se } u_{Med}(A, B) < 0 \\ 2 \sum_{i=0}^{\frac{n_A n_B}{2}-1} (-1)^i I_{u_{Med}(A, B)} \left(\frac{n_A n_B}{2} + i + 1, \frac{n_A n_B}{2} - i \right) & \text{se } 0 \leq u_{Med}(A, B) \leq 0.5 \\ 2 \sum_{i=0}^{\frac{n_A n_B}{2}-1} (-1)^i I_{u_{Med}(A, B)} \left(\frac{n_A n_B}{2} + i + 1, \frac{n_A n_B}{2} - i \right) + (-1)^{\frac{n_A n_B}{2}} (2u_{Med}(A, B) - 1)^{n_A n_B} & \text{se } 0.5 < u_{Med}(A, B) \leq 1 \\ 1 & \text{se } u_{Med}(A, B) > 1 \end{cases}$$

sendo

$$I_{u_{Med}(A, B)}(a, b) \equiv \frac{B_{u_{Med}(A, B)}(a, b)}{B(a, b)} \equiv \frac{1}{B(a, b)} \int_0^{u_{Med}(A, B)} t^{a-1} (1-t)^{b-1} dt, \quad a, b > 0,$$

com valores limite $I_0(a, b) = 0$ e $I_1(a, b) = 1$, para a e b inteiros.

De seguida, apresenta-se a dedução das expressões do valor esperado e da variância da função de distribuição da mediana amostral de um conjunto de dimensão par, as quais se obtêm recorrendo a alguns cálculos algébricos. Para o que se segue, por simplicidade de escrita, usa-se X em vez de U_{med} .

Considere-se a expressão da função densidade da estatística mediana de um conjunto amostral de cardinal par, $N = 2r$, com r inteiro positivo (Sousa [40]),

$$f(x) = \begin{cases} 2 \sum_{i=0}^{r-1} (-1)^i \frac{x^{r+i} (1-x)^{r-i-1}}{B(r+i+1, r-i)} & \text{se } 0 \leq x \leq 0.5 \\ 2 \sum_{i=0}^{r-1} (-1)^i \frac{x^{r+i} (1-x)^{r-i-1}}{B(r+i+1, r-i)} + 4r(-1)^r (2x-1)^{2r-1} & \text{se } 0.5 < x \leq 1 \\ 0 & \text{c. c.} \end{cases}$$

Tem-se

$$\begin{aligned}
 E(X) &= \int_0^{0.5} \sum_{i=0}^{r-1} (-1)^i \frac{2}{B(r+i+1, r-i)} x^{r+i+1} (1-x)^{r-i-1} dx + \\
 &\quad + \int_{0.5}^1 \left[\sum_{i=0}^{r-1} (-1)^i \frac{2}{B(r+i+1, r-i)} x^{r+i+1} (1-x)^{r-i-1} + 4r(-1)^r x(2x-1)^{2r-1} \right] dx \\
 &= \sum_{i=0}^{r-1} (-1)^i \frac{2}{B(r+i+1, r-i)} \int_0^1 x^{r+i+1} (1-x)^{r-i-1} dx + \\
 &\quad + \int_{0.5}^1 4r(-1)^r x(2x-1)^{2r-1} dx.
 \end{aligned}$$

Considerando a mudança de variável $y = 2x - 1$, obtém-se

$$\begin{aligned}
 E(X) &= \sum_{i=0}^{r-1} (-1)^i \frac{2}{B(r+i+1, r-i)} \times B(r+i+2, r-i) + \int_0^1 (-1)^r r(y^{2r} + y^{2r-1}) dy \\
 &= \sum_{i=0}^{r-1} 2(-1)^i \frac{2r!(r+i+1)!(r-i-1)!}{(2r+1)!(r+i)!(r-i-1)!} + (-1)^r r \left[\frac{y^{2r+1}}{2r+1} + \frac{y^{2r}}{2r} \right]_0^1 \\
 &= \sum_{i=0}^{r-1} 2(-1)^i \frac{r+i+1}{2r+1} + (-1)^r r \left(\frac{1}{2r+1} + \frac{1}{2r} \right) \\
 &= \frac{2r+2}{2r+1} \sum_{i=0}^{r-1} (-1)^i + \frac{2}{2r+1} \sum_{i=0}^{r-1} (-1)^i i + (-1)^r r \left(\frac{1}{2r+1} + \frac{1}{2r} \right).
 \end{aligned}$$

Como $\sum_{i=0}^{r-1} (-1)^i i = -\frac{(-1)^r r}{2} + \frac{(-1)^r}{4} - \frac{1}{4}$, tem-se

$$E(X) = \frac{2r+2}{2r+1} \sum_{i=0}^{r-1} (-1)^i + \frac{2}{2r+1} \left[-\frac{(-1)^r r}{2} + \frac{(-1)^r}{4} - \frac{1}{4} \right] + (-1)^r r \left(\frac{1}{2r+1} + \frac{1}{2r} \right).$$

Sendo $\sum_{i=0}^{r-1} (-1)^i = \begin{cases} 0 & \text{se } r \text{ par} \\ 1 & \text{se } r \text{ impar} \end{cases}$ conclui-se que

$$E(X) = \frac{1}{2}.$$

$$\begin{aligned}
 E(X^2) &= \int_0^{0.5} \sum_{i=0}^{r-1} (-1)^i \frac{2}{B(r+i+1, r-i)} x^{r+i+2} (1-x)^{r-i-1} dx + \\
 &\quad + \int_{0.5}^1 \sum_{i=0}^{r-1} (-1)^i \frac{2}{B(r+i+1, r-i)} x^{r+i+2} (1-x)^{r-i-1} dx + \int_{0.5}^1 4r(-1)^r x^2(2x-1)^{2r-1} dx.
 \end{aligned}$$

Considerando novamente a mudança de variável $y = 2x - 1$, tem-se

$$\begin{aligned}
 E(X^2) &= \sum_{i=0}^{r-1} \frac{2(-1)^i}{B(r+i+1, r-i)} \int_0^1 x^{r+i+2} (1-x)^{r-i-1} dx + \int_0^1 2r(-1)^r \left(\frac{y+1}{2}\right)^2 y^{2r-1} dy \\
 &= \sum_{i=0}^{r-1} \frac{2(-1)^i}{B(r+i+1, r-i)} \int_0^1 x^{r+i+2} (1-x)^{r-i-1} dx + \frac{(-1)^r r}{2} \int_0^1 (y^2 + 2y + 1) y^{2r-1} dy \\
 &= \sum_{i=0}^{r-1} \frac{2(-1)^i}{B(r+i+1, r-i)} \times B(r+i+3, r-i) + \frac{(-1)^r r}{2} \int_0^1 (y^{2r+1} + 2y^{2r} + y^{2r-1}) dy \\
 &= \sum_{i=0}^{r-1} 2(-1)^i \frac{(r+i+2)(r+i+1)}{(2r+2)(2r+1)} + (-1)^r \left(\frac{r}{4r+4} + \frac{r}{2r+1} + \frac{1}{4} \right) \\
 &= \frac{r^2 + 3r + 2}{(r+1)(2r+1)} \sum_{i=0}^{r-1} (-1)^i + \frac{2r+3}{(r+1)(2r+1)} \sum_{i=0}^{r-1} (-1)^i i + \\
 &\quad + \frac{1}{(r+1)(2r+1)} \sum_{i=0}^{r-1} (-1)^i i^2 + (-1)^r \left(\frac{r}{4r+4} + \frac{r}{2r+1} + \frac{1}{4} \right).
 \end{aligned}$$

Dado que $\sum_{i=0}^{r-1} (-1)^i i = -\frac{(-1)^r r}{2} + \frac{(-1)^r}{4} - \frac{1}{4}$ e $\sum_{i=0}^{r-1} (-1)^i i^2 = \frac{(-1)^r r}{2} - \frac{(-1)^r r^2}{2}$ tem-se

$$E(X^2) = \frac{2r^2 + 4r + 1}{4(r+1)(2r+1)}.$$

Logo, a variância da estatística mediana amostral de dimensão par $N = 2r$ é dada por

$$Var(X) = \frac{r}{4(r+1)(2r+1)} = \frac{N}{(2N+2)(2N+4)}.$$

Para a programação da função de distribuição exacta da mediana, em caso de dimensão amostral par, recorre-se ao algoritmo de cálculo da função beta incompleta. O algoritmo implementado foi obtido de Sousa [40]. O cálculo de $B(m, n)$ e $B_x(m, n)$ ¹ foi efectuado por recurso ao seguinte algoritmo:

- Cálculo de $B(m, n)$:

$$b_i = \frac{i}{m+i} b_{i-1}, \quad i = 1, \dots, n-1,$$

$$\text{para } b_{n-1} = B(m, n) \text{ e } b_0 = \frac{1}{m}.$$

¹Abusivamente é usada a notação $B(m, n)$ e $B_x(m, n)$ por ser a notação usual para esta função.

- Cálculo de $B_x(m, n)$:

$$\begin{aligned}
 - B_x(m, n) &= \sum_{i=0}^{n-1} \frac{(m-1)!(n-1)!}{(m+i)!(n-i-1)!} x^{m+i} (1-x)^{n-i-1}; \\
 - c_0 &= \frac{1}{m} x^m (1-x)^{n-1}; \\
 - c_i &= \frac{(n-i)x}{(m+i)(1-x)} c_{i-1}, \text{ com } i = 1, \dots, n-1; \\
 \text{vindo } \sum_{i=0}^{n-1} c_i &= B_x(m, n).
 \end{aligned}$$

Transpondo para a Classificação Não Hierárquica, se n_A é um número par, durante o processo de afectação, x será atribuído à classe A se é máximo o valor da função objectivo

$$\Gamma_{NHVMED}(\{x\}, A) =$$

$$= \begin{cases} 0, & \text{se } z < 0 \\ 2 \sum_{i=0}^{\frac{n_A}{2}-1} (-1)^i I_z \left(\frac{n_A}{2} + i + 1, \frac{n_A}{2} - i \right), & \text{se } 0 \leq z \leq 0.5 \\ 2 \sum_{i=0}^{\frac{n_A}{2}-1} (-1)^i I_z \left(\frac{n_A}{2} + i + 1, \frac{n_A}{2} - i \right) + (-1)^{\frac{n_A}{2}} (2z-1)^{n_A}, & \text{se } 0.5 < z \leq 1 \\ 1, & \text{se } z > 1 \end{cases} \quad (3.1.13)$$

para $z = u_{Med}(\{x\}, A) = \text{mediana}\{\gamma(x, a) : a \in A\}$.

3.1.4 Método NHVC

Dadas duas classes A e B , de cardinais n_A e n_B , respectivamente, admita-se de novo que o conjunto $\{\gamma(a, b) : (a, b) \in A \times B\}$ constitui uma amostra independente e identicamente distribuída, de uma distribuição $U(0, 1)$.

No critério AVC, de Classificação Hierárquica, considera-se como estatística o centro amostral U_C , dado por:

$$U_C = \frac{\max \gamma(a, b) + \min \gamma(a, b)}{2}, (a, b) \in A \times B,^2$$

²Para simplificação de escrita usa-se $\max \gamma(a, b)$ (respectivamente, $\min \gamma(a, b)$) para representar $\max \{\gamma(a, b) : (a, b) \in A \times B\}$ (respectivamente, $\min \{\gamma(a, b) : (a, b) \in A \times B\}$).

cua função de distribuição, na hipótese de referência admitida é (Sousa [40]):

$$\Gamma_{VC}(A, B) = \begin{cases} 0 & \text{se } u_C(A, B) < 0 \\ 2^{n_A n_B - 1} u_C(A, B)^{n_A n_B} & \text{se } 0 \leq u_C(A, B) \leq 0.5 \\ 1 - 2^{n_A n_B - 1} (1 - u_C(A, B))^{n_A n_B} & \text{se } 0.5 < u_C(A, B) \leq 1 \\ 1 & \text{se } u_C(A, B) > 1 \end{cases}$$

No caso da Classificação Não Hierárquica, tendo uma das classes cardinal igual a um, cada elemento x , será afectado à classe que maximize o valor da função de distribuição

$$\Gamma_{NHVC}(\{x\}, A) = \begin{cases} 0 & \text{se } u_C(\{x\}, A) < 0 \\ 2^{n_A - 1} u_C(\{x\}, A)^{n_A} & \text{se } 0 \leq u_C(\{x\}, A) \leq 0.5 \\ 1 - 2^{n_A - 1} (1 - u_C(\{x\}, A))^{n_A} & \text{se } 0.5 < u_C(\{x\}, A) \leq 1 \\ 1 & \text{se } u_C(\{x\}, A) > 1 \end{cases} \quad (3.1.14)$$

em que $u_C(\{x\}, A) = \frac{\max \gamma(x, a) + \min \gamma(x, a)}{2}$, com $a \in A$.

3.2 Métodos Directos: Abordagem Clássica

Como foi anteriormente referido, o tema deste trabalho assenta numa abordagem probabilística em métodos de Classificação Não Hierárquica, uma vez que os métodos desenvolvidos recorrem a noções probabilísticas na definição de ambas as funções de comparação, entre elementos e entre classes. Nesta secção apresentam-se métodos que são denominados de **Métodos Directos**, os quais se caracterizam pela função de afectação de um elemento a uma classe ser dada pelo valor de uma estatística das semelhanças VL entre pares de elementos.

Em Classificação Não Hierárquica, o uso directo de estatísticas amostrais de tendência central, como a média aritmética, a média geométrica, o centro e a mediana, como critério de agregação dará lugar aos métodos **NHDMEAN** (Non Hierarchical Direct MEAN), **NHDGM** (Non Hierarchical Direct Geometric Mean), **NHDC** (Non Hierarchical Direct Center) e **NHDMED** (Non Hierarchical Direct MEDian).

Com a comparação dos resultados obtidos pelos métodos directos com aqueles que

se baseiam na metodologia VL, pretende-se avaliar as vantagens do uso da função de distribuição de estatísticas amostrais na definição das funções de afectação.

3.2.1 Método NHDMEAN

Neste critério, a função de comparação entre duas classes, A e B , de cardinais n_A e n_B , respectivamente, baseia-se na média aritmética dos valores do índice de comparação entre elementos das classes a comparar, $\gamma(a, b)$, ou seja,

$$\Gamma_{MEAN}(A, B) = \frac{\sum_{(a,b) \in A \times B} \gamma(a, b)}{n_A n_B}.$$

Em Classificação Não Hierárquica, uma das classes é singular logo, o critério de agregação no método NHDMEAN é dado por

$$\Gamma_{NHDMEAN}(\{x\}, A) = \frac{\sum_{a \in A} \gamma(x, a)}{n_A}.$$

3.2.2 Método NHDGM

Com este critério, a função de comparação entre duas classes A e B , é a média geométrica das semelhanças entre os pares de elementos de cada uma das classes,

$$n_A n_B \sqrt[n_A n_B]{\prod_{(a,b) \in A \times B} \gamma(a, b)}. \quad (3.2.1)$$

Aplicando logaritmos neperianos a (3.2.1) e considerando o seu simétrico, obtém-se

$$\Gamma_{GM}(A, B) = \frac{\sum_{(a,b) \in A \times B} -\ln \gamma(a, b)}{n_A n_B}.$$

Considerando B uma classe com um único elemento, o elemento x será atribuído à classe A que minimize a função objectivo

$$\Gamma_{NHDGM}(\{x\}, A) = \frac{\sum_{a \in A} -\ln \gamma(x, a)}{n_A}.$$

3.2.3 Método NHD MED

Baseando-se o método NHD MED na estatística mediana amostral, a função de comparação entre duas classes A e B , é dada por

$$\Gamma_{MED}(A, B) = \text{mediana}\{\gamma(a, b) : (a, b) \in A \times B\}.$$

Em Classificação Não Hierárquica, afecta-se x à classe A , que maximiza a função objectivo

$$\Gamma_{NHD MED}(\{x\}, A) = \text{mediana}\{\gamma(x, a) : a \in A\}.$$

3.2.4 Método NHDC

Neste método a comparação entre as classes A e B , é dada pelo valor da expressão

$$\Gamma_C(A, B) = \frac{\min \gamma(a, b) + \max \gamma(a, b)}{2}, \quad (a, b) \in A \times B.$$

Naturalmente se uma das classes tem cardinal um, a função objectivo do método NHDC é dada por

$$\Gamma_{NHDC}(\{x\}, A) = \frac{\min \gamma(x, a) + \max \gamma(x, a)}{2}, \quad a \in A.$$

Os algoritmos dos métodos directos expostos compreendem os mesmos passos do algoritmo básico de um método do tipo centróide conforme é apresentado na Secção 3.4, diferindo apenas na função objectivo que é usada, e na medida de comparação entre pares de elementos.

3.3 Critério *Wise-Choice*

Um dos aspectos mais importantes em métodos não hierárquicos é a escolha/determinação dos centros iniciais. O critério *wise-choice* (Nicolau [31]), foi o que se manifestou mais eficaz na rápida convergência do algoritmo em estudos realizados por Brito [9] e Nicolau e Brito [33], quando usado como critério de inicialização do método NHMEAN.

Assumindo que γ é um índice de comparação entre elementos de tipo semelhança, considere-se o conjunto dos valores de semelhanças entre todos os pares de indivíduos de E , $\{\gamma(I_i, I_\ell) : I_i, I_\ell \in E\}$.

Para um valor fixo e natural de L , define-se a **L-vizinhança** de I_i , $V_L(I_i)$, como

$$V_L(I_i) = \{I_\ell \in E : I_\ell \text{ é um dos } L \text{ elementos mais próximos de } I_i\}$$

isto é, $V_L(I_i)$ é o conjunto dos L elementos de E mais próximos de I_i .

Como $\gamma(I_i, I_i) = \gamma_{\max}$, então $I_i \in V_L(I_i)$. Obviamente, $V_L(I_i)$ terá apenas $L - 1$ elementos distintos de I_i .

Seja $S_L(I_i)$ a média das semelhanças dos $L - 1$ elementos mais próximos de I_i

$$S_L(I_i) = \frac{\sum_{I_\ell \in V_L(I_i) \setminus \{I_i\}} \gamma(I_i, I_\ell)}{L - 1}, I_i \in E. \quad (3.3.1)$$

Na procura dos k centros, toma-se para primeiro centro c_1 , o ponto que maximiza $S_L(I_i)$, $\forall I_i \in E$, isto é,

$$c_1 \equiv I_\ell \in E : S_L(I_\ell) = \max\{S_L(I_i) : I_i \in E\}. \quad (3.3.2)$$

Os restantes $k - 1$ centros, c_j , com $j = 2, \dots, k$, são os elementos que satisfazem a condição

$$c_j \equiv I_\ell \in E :$$

$$\min_{1 \leq m \leq j-1} \frac{S_L(I_\ell)}{\gamma(I_\ell, c_m)} = \max \left\{ \min_{1 \leq m \leq j-1} \frac{S_L(I_i)}{\gamma(I_i, c_m)} : I_i \in E \setminus \{c_1, \dots, c_{j-1}\} \right\} \quad (3.3.3)$$

De entre todos os elementos de E , c_j é o elemento mais distante dos centros já determinados, e localizado numa vizinhança densa.

Sendo γ uma função de comparação do tipo dissemelhança e $S_L(I_i)$ dado por (3.3.1), o primeiro centro c_1 será então o elemento I_ℓ que satisfaz a condição

$$c_1 \equiv I_\ell \in E : S_L(I_\ell) = \min\{S_L(I_i) : I_i \in E\}.$$

A expressão que fornece os centros c_j , com $j = 2, \dots, k$, é dada por

$$c_j \equiv I_\ell \in E : \max_{1 \leq m \leq j-1} \frac{S_L(I_\ell)}{\gamma(I_\ell, c_m)} = \min \left\{ \max_{1 \leq m \leq j-1} \frac{S_L(I_i)}{\gamma(I_i, c_m)} : I_i \in E \setminus \{c_1, \dots, c_{j-1}\} \right\}.$$

Observando a expressão (3.3.1), que se trata da média aritmética das semelhanças entre um elemento e os restantes elementos pertencentes à sua L -vizinhança, é evidente a relação com a estatística subjacente ao método NIIMEAN. Esta observação terá como consequência natural a reformulação das expressões (3.3.1) e (3.3.3) de acordo com a estatística usada em cada um dos métodos propostos no trabalho, dando origem às seguintes adaptações:

- No método NHVC, o primeiro pólo de agregação c_1 , é o elemento que maximiza

$$C_L(I_i) = \frac{\min_{I_\ell \in V_L(I_i) \setminus \{I_i\}} \gamma(I_i, I_\ell) + \max_{I_\ell \in V_L(I_i) \setminus \{I_i\}} \gamma(I_i, I_\ell)}{2}, I_i \in E.$$

Os restantes centros obtêm-se da relação

$$c_j \equiv I_\ell \in E :$$

$$\min_{1 \leq m \leq j-1} \frac{C_L(I_\ell)}{\gamma(I_\ell, c_m)} = \max \left\{ \min_{1 \leq m \leq j-1} \frac{C_L(I_i)}{\gamma(I_i, c_m)} : I_i \in E \setminus \{c_1, \dots, c_{j-1}\} \right\} \quad (3.3.4)$$

- No método NIIVGM, a expressão (3.3.1) é substituída por

$$MG_L(I_i) = \frac{\sum_{I_\ell \in V_L(I_i) \setminus \{I_i\}} \ln \gamma(I_i, I_\ell)}{L - 1}. \quad (3.3.5)$$

Os restantes centros são os elementos do conjunto que satisfazem uma condição análoga a (3.3.3) substituindo o numerador pelo valor $MG_L(I_\ell)$ (respectivamente, $MG_L(I_i)$).

• No método NHVMED tem-se:

- se L par, o primeiro centro é o elemento $I_i \in E$ que maximiza o valor da estatística de ordem $\frac{L}{2}$,

$$\begin{aligned} MED_L(I_i) &= \text{mediana}\{\gamma(I_i, I_\ell) : I_\ell \in V_L(I_i) \setminus \{I_i\}\} \\ &= \gamma_{(\frac{L}{2}:L-1)}(I_i, I_\ell) \end{aligned}$$

- se L ímpar, o primeiro centro é o elemento $I_i \in E$ que maximiza a expressão

$$\begin{aligned} MED_L(I_i) &= \text{mediana}\{\gamma(I_i, I_\ell) : I_\ell \in V_L(I_i) \setminus \{I_i\}\} \\ &= \frac{\gamma_{(\frac{L-1}{2}:L-1)}(I_i, I_\ell) + \gamma_{(\frac{L+1}{2}:L-1)}(I_i, I_\ell)}{2} \end{aligned}$$

Em ambos os casos os centros c_j , com $j = 2, \dots, k$, são os pontos que obedecem a uma condição semelhante à que figura em (3.3.3), substituindo o numerador $S_L(I_i)$ por $MED_L(I_i)$.

O mesmo tipo de adaptações que acabam de ser expostas para o critério de inicialização *wise-choice*, foram efectuadas no critério que acenta na escolha dos k pontos mais afastados na amostra e mais afastados entre si, abordado na Secção 2.1.

No critério NIIVC, a expressão (2.1.1) é substituída pelo centro amostral das semelhanças relativamente a todos os restantes indivíduos da classe:

$$C(I_i) = \frac{\max \gamma(I_i, I_\ell) + \min \gamma(I_i, I_\ell)}{2}, I_\ell \in E \setminus \{I_i\}. \quad (3.3.6)$$

O primeiro centro c_1 é o elemento que minimiza (3.3.6). Os restantes elementos obedecem a uma condição semelhante à de (2.1.3), substituindo $S(I_i)$ por $C(I_i)$.

No critério NIIVGM, a expressão em (2.1.1) é substituída por

$$MG(I_i) = \frac{1}{n-1} \sum_{I_\ell \in E \setminus \{I_i\}} \ln \gamma(I_i, I_\ell).$$

No critério NHMED, a expressão (2.1.1) substitui-se por

$$MED(I_i) = \text{mediana}\{\gamma(I_i, I_\ell) : I_\ell \in E \setminus \{I_i\}\}.$$

Também nos métodos NHVGM e NHMED, os centros são os pontos que satisfazem (2.1.2) e (2.1.3), substituindo $S(I_i)$ por $MG(I_i)$ e $MED(I_i)$, respectivamente.

3.4 Algoritmo Básico

Baseado nos principais passos de um método centróide, o algoritmo de Classificação Não Hierárquica que se segue opera iterativamente sobre um conjunto de dados construindo sucessivamente a partição final. A classificação é determinada em função da dissimilaridade dos elementos aos centros das classes, que funcionam como pólos de agregação; no entanto, distingue-se dos restantes pelo uso da metodologia VL. Assim, de um modo geral, os passos do algoritmo são:

1. Escolha dos centros iniciais.
2. Afectação de cada elemento ao centro mais próximo.
3. Actualização dos centros das classes da partição obtida.
4. Reclassificação dos elementos.

Repetem-se os passos 3 e 4 até convergência.

Com recurso à estatística média aritmética, a metodologia aqui desenvolvida foi já explorada por Brito [9], no método NHMEAN. As etapas integrantes do algoritmo implementado são comuns aos métodos NIIMEAN, NIIVC, NHVGM e NHVMED, pelo que se optou por fazer a apresentação em comum, especificando as particularidades de cada método, quando necessário.

Mais uma vez, considere-se um conjunto E de cardinal n e os valores do índice de comparação entre os elementos de E ,

$$\{\gamma(I_i, I_\ell) : I_i, I_\ell \in E\}.$$

Fixado o número máximo de classes pretendido, k , e o número máximo de iterações permitido, o algoritmo engloba as seguintes etapas:

Etapla 0: Inicialização (Determinação dos centros iniciais)

A escolha dos centros iniciais, que funcionam como pólos de atracção, pode ser feita de entre algumas das opções já referidas na Secção 2.1. Em particular, no programa elaborado encontram-se disponíveis os seguintes critérios:

- a) Escolha dos k primeiros elementos da amostra.
- b) Escolha aleatória de k elementos.
- c) Escolha dos elementos de ordem $\frac{j * n}{k}$, $j = 1, \dots, k$ (no caso dos valores não serem inteiros consideram-se os valores característicos).
- d) Escolha subjectiva de k elementos.
- e) Escolha dos k elementos, simultaneamente, mais afastados na amostra e mais afastados entre si.
- f) Critério de Lerman.
- g) Critério *wise-choice*.

Etapla 1: Agregação (Classificação dos restantes elementos)

Para o que se segue a notação usada é a seguinte:

- A_j^h representa a j -ésima classe, de cardinal n_j^h , obtida no h -ésimo passo desta etapa.
- $u(\{I_i\}, A_j^h)$ representa o valor da estatística amostral sobre o conjunto dos valores do índice de comparação entre I_i e os indivíduos I_ℓ pertencentes à classe A_j^h ,

$$\left\{ \gamma(I_i, I_\ell) : I_\ell \in A_j^h \right\}. \quad (3.4.1)$$

Em particular, denota-se por $u_M(\{I_i\}, A_j^h)$, $u_C(\{I_i\}, A_j^h)$, $u_{MED}(\{I_i\}, A_j^h)$, $u_{MG}(\{I_i\}, A_j^h)$, a média aritmética, o centro, a mediana e a média geométrica, do conjunto de valores em (3.4.1), respectivamente.

- $\Gamma(\{I_i\}, A_j^h)$ representa o valor da função de distribuição da estatística usada, no ponto $u(\{I_i\}, A_j^h)$.

Escolhidos os k centros iniciais c_j , com $j = 1, \dots, k$, elementos do conjunto a classificar, procede-se à reordenação do conjunto dos indivíduos e consequentemente, à reordenação da matriz dos valores do índice de comparação entre pares de elementos, de forma a que c_j , $j = 1, \dots, k$, sejam os k primeiros elementos de E . Deste modo, I_{k+h} , $h = 1, \dots, n - k$, é o elemento a classificar no h -ésimo passo.

Seja c_j o elemento representante da classe A_j^0 . Então, na etapa 0 do processo tem-se $A_j^0 = \{c_j\}$, $n_j^0 = 1$, para $j = 1, \dots, k$.

1.1. Afecção de I_{k+h}

Para os elementos, I_{k+h} , com $h = 1, \dots, n - k$, a afecção realiza-se do seguinte modo: o algoritmo afecta o elemento I_{k+h} à classe $A_j^{h'}$, $h' = 0, \dots, n - k - 1$, para a qual é máximo o valor de $\Gamma(\{I_{k+h}\}, A_j^{h'})$. A classe destino de I_{k+h} , $A_j^{h'}$, é tal que

$$\Gamma(\{I_{k+h}\}, A_j^{h'}) = \max\{\Gamma(\{I_{k+h}\}, A_j^{h'}) : j = 1, \dots, k\}. \quad (3.4.2)$$

No caso do elemento de ordem $k + 1$, afectar o elemento à classe que maximize $\Gamma(\{I_{k+1}\}, A_j^0)$ é equivalente a maximizar $\gamma(I_{k+1}, c_j)$, uma vez que no momento da afecção deste elemento todas as classes são constituídas por um único elemento.

1.1.1. Após a afecção do elemento I_{k+h} , $h = 1, \dots, n - k$, à classe $A_j^{h'}$, procede-se às seguintes actualizações:

- $A_j^h = A_j^{h'} \cup \{I_{k+h}\}$, $n_j^h = n_j^{h'} + 1$;
- $A_j^h = A_j^{h'}$, $n_j^h = n_j^{h'}$, $j = 1, \dots, k$, $j \neq j'$.

1.1.2. Nesta fase do algoritmo é feita a actualização do valor da estatística usada e do seu valor na respectiva função de distribuição. Naturalmente, surge o seguinte desdobramento:

I. Método NHMEAN

Neste método a estatística base é a média aritmética amostral, logo

$$i) \ u_M(I_\ell, A_{j'}^h) = \frac{n_{j'}^{h'} u_M(I_\ell, A_{j'}^{h'}) + \gamma(I_\ell, I_{k+h})}{n_{j'}^h}, \ell = k + h + 1, \dots, n.$$

$$ii) \ u_M(\{I_\ell\}, A_j^h) = u_M(\{I_\ell\}, A_{j'}^h), \ell = k + h + 1, \dots, n, j = 1, \dots, k, j \neq j'.$$

iii) Para $\ell = k + h + 1$, o valor de $\Gamma_{NHMEAN}(\{I_\ell\}, A_j^h)$ é dado pelo valor de (3.1.2) em $u_M(\{I_\ell\}, A_j^h)$.

II. Método NHVC

Tendo por base a estatística centro amostral, tem-se

$$i) \ u_C(\{I_\ell\}, A_{j'}^h) = \frac{\max \gamma(I_\ell, I_i) + \min \gamma(I_\ell, I_i)}{2}, I_i \in A_{j'}^{h'}, \ell = k + h + 1, \dots, n.$$

$$ii) \ u_C(\{I_\ell\}, A_j^h) = u_C(\{I_\ell\}, A_{j'}^h), \ell = k + h + 1, \dots, n, j \neq j'.$$

iii) Para $\ell = k + h + 1$, $\Gamma_{NHVC}(\{I_\ell\}, A_j^h)$ é o valor da função (3.1.14) em $u_C(\{I_\ell\}, A_{j'}^h)$.

III. Método NHVGM

A estatística usada é a média geométrica amostral, logo

$$i) \ u_{MG}(\{I_\ell\}, A_{j'}^h) = \frac{n_{j'}^{h'} u_{MG}(I_\ell, A_{j'}^{h'}) - \ln \gamma(I_\ell, I_{k+h})}{n_{j'}^h}, \ell = k + h + 1, \dots, n.$$

$$ii) \ u_{MG}(\{I_\ell\}, A_j^h) = u_{MG}(\{I_\ell\}, A_{j'}^h), \ell = k + h + 1, \dots, n, j \neq j'.$$

iii) Para $\ell = k + h + 1$, $\Gamma_{NHVGM}(\{I_\ell\}, A_j^h)$ obtém-se substituindo $u_{MG}(\{I_\ell\}, A_{j'}^h)$ em (3.1.8).

IV. Método NHVMED

Neste método a estatística base é a mediana amostral, logo

$$i) \ u_{Med}(\{I_\ell\}, A_{j'}^h) = \text{mediana}\{\gamma(I_\ell, I_i) : I_i \in A_{j'}^{h'}\}, \ell = k + h + 1, \dots, n.$$

$$ii) \ u_{Med}(\{I_\ell\}, A_j^h) = u_{Med}(\{I_\ell\}, A_{j'}^h), \ell = k + h + 1, \dots, n, j \neq j'.$$

iii) Para $\ell = k + h + 1$:

- Se $n_{j'}^h$ é ímpar $\Gamma_{NHVMED}(\{I_\ell\}, A_j^h)$ é o valor da função em (3.1.12) no ponto $u_{Med}(\{I_\ell\}, A_{j'}^h)$.

- Se n_j^h é par, $\Gamma_{NHVMED}(\{I_\ell\}, A_j^h)$ é o valor da função em (3.1.13) no ponto $u_{Med}(\{I_\ell\}, A_j^h)$.

1.2.3. Tome-se $h' = h$ e $h = h' + 1$.

Esta etapa 1. termina com a afectação de todos os elementos, o que envolve $n - k$ classificações.

Etapa 2: Definição dos novos centros das classes obtidas

Considerando U a variável aleatória que representa a estatística usada e Γ_U a sua função de distribuição, então para cada uma das classes A_j^{n-k} da partição obtida, toma-se para centro o ponto $c_j \in A_j^{n-k}$ que maximiza

$$\Gamma_U(\{c_j\}, A_j^{n-k}) = P(U \leq u(\{c_j\}, A_j^{n-k}))$$

na hipótese de referência, o que é equivalente a determinar o elemento da classe que maximiza o valor da estatística do índice de comparação entre elementos relativamente aos restantes elementos das classes. Ora, se n_j^{n-k} é o cardinal de A_j^{n-k} , com $j = 1, \dots, k$, a função de distribuição de Γ_U é a mesma para qualquer um dos elementos desta classe, dependendo apenas do parâmetro n_j^{n-k} , que é igual em qualquer dos casos, assim,

$$c_j \equiv I_i \in A_j^{n-k} : u(\{I_i\}, A_j^{n-k} \setminus \{I_i\}) = \max \left\{ u(\{I_\ell\}, A_j^{n-k} \setminus \{I_\ell\}) : I_\ell \in A_j^{n-k} \right\}.$$

I. Método NHMEAN

Pelo que foi observado no parágrafo anterior, no método NHMEAN o centro da classe A_j^{n-k} é o elemento que maximiza a semelhança média relativamente aos restantes elementos de A_j^{n-k} , ou seja,

$$\begin{aligned} c_j &\equiv I_i : \frac{1}{n_j^{n-k} - 1} \sum_{I_\ell \in A_j^{n-k} \setminus \{I_i\}} \gamma(I_i, I_\ell) \\ &= \max \left\{ \frac{1}{n_j^{n-k} - 1} \sum_{I_\ell \in A_j^{n-k} \setminus \{I_{i'}\}} \gamma(I_{i'}, I_\ell) : I_{i'} \in A_j^{n-k} \right\}. \end{aligned}$$

II. Método NHVC

Dada a classe A_j^{n-k} , o centro c_j é o ponto que maximiza o centro das semelhanças relativamente aos restantes elementos de A_j^{n-k} :

$$\begin{aligned} c_j &\equiv I_i : \frac{\max_{I_\ell \in A_j^{n-k} \setminus \{I_i\}} \gamma(I_i, I_\ell) + \min_{I_\ell \in A_j^{n-k} \setminus \{I_i\}} \gamma(I_i, I_\ell)}{2} = \\ &= \max \left\{ \frac{\max_{I_\ell \in A_j^{n-k} \setminus \{I_{i'}\}} \gamma(I_{i'}, I_\ell) + \min_{I_\ell \in A_j^{n-k} \setminus \{I_{i'}\}} \gamma(I_{i'}, I_\ell)}{2} : I_{i'} \in A_j^{n-k} \right\}. \end{aligned}$$

III. Método NHVGM

O centro da classe A_j^{n-k} é o ponto c_j que maximiza a média geométrica das semelhanças relativamente aos restantes elementos da classe, ou simplesmente, este critério corresponde a determinar o elemento da classe cuja média do logaritmo neperiano das semelhanças tem a menor probabilidade de ser ultrapassada, na hipótese de referência. Deste modo,

$$\begin{aligned} c_j &\equiv I_i : \frac{1}{n_j^{n-k} - 1} \sum_{I_\ell \in A_j^{n-k} \setminus \{I_i\}} \ln \gamma(I_i, I_\ell) = \\ &= \max \left\{ \frac{1}{n_j^{n-k} - 1} \sum_{I_\ell \in A_j^{n-k} \setminus \{I_{i'}\}} \ln \gamma(I_{i'}, I_\ell) : I_{i'} \in A_j^{n-k} \right\}. \end{aligned}$$

Considerando o simétrico do logaritmo neperiano das semelhanças tem-se

$$\begin{aligned} c_j &\equiv I_i : \frac{1}{n_j^{n-k} - 1} \sum_{I_\ell \in A_j^{n-k} \setminus \{I_i\}} -\ln \gamma(I_i, I_\ell) = \\ &= \min \left\{ \frac{1}{n_j^{n-k} - 1} \sum_{I_\ell \in A_j^{n-k} \setminus \{I_{i'}\}} -\ln \gamma(I_{i'}, I_\ell) : I_{i'} \in A_j^{n-k} \right\}. \end{aligned}$$

IV. Método NHVMED

Dada a classe A_j^{n-k} , toma-se para centro o ponto que maximiza a mediana das semelhanças relativamente aos restantes $n_j^{n-k} - 1$ elementos desta classe. Assim, ordenando o conjunto

$$\{\gamma(c_j, I_i) : I_i \in A_j^{n-k} \setminus \{c_j\}\}$$

tem-se

i) Se n_j^{n-k} é par

$$\begin{aligned} c_j &\equiv I_i : \gamma\left(\frac{n_j^{n-k}}{2} : n_j^{n-k}-1\right)(I_i, I_\ell) = \\ &= \max \left\{ \gamma\left(\frac{n_j^{n-k}}{2} : n_j^{n-k}-1\right)(I_{i'}, I_\ell) : I_{i'} \in A_j^{n-k} \right\}, I_\ell \in A_j^{n-k} \setminus \{I_i\}. \end{aligned}$$

ii) Se n_j^{n-k} é ímpar

$$\begin{aligned} c_j &\equiv I_i : \frac{\gamma\left(\frac{n_j^{n-k}-1}{2} : n_j^{n-k}-1\right)(I_i, I_\ell) + \gamma\left(\frac{n_j^{n-k}+1}{2} : n_j^{n-k}-1\right)(I_i, I_\ell)}{2} = \\ &= \max \left\{ \frac{\gamma\left(\frac{n_j^{n-k}-1}{2} : n_j^{n-k}-1\right)(I_{i'}, I_\ell) + \gamma\left(\frac{n_j^{n-k}+1}{2} : n_j^{n-k}-1\right)(I_{i'}, I_\ell)}{2} : I_{i'} \in A_j^{n-k} \right\}, \\ &I_\ell \in A_j^{n-k} \setminus \{I_i\} \end{aligned}$$

Etapa 3: Reiteração

Considere-se P^r a partição obtida na r -ésima iteração, isto é,

$$P^r = \{A_j^{n-k} : j = 1, \dots, k\}.$$

Para $r > 1$, comparam-se as partições P^{r-1} e P^r . Se a convergência fôr atingida, o algoritmo pára, caso contrário, armazena-se P^r . Neste último caso, definidos os centros das classes da r -ésima partição, dá-se início a uma nova iteração com a etapa 1. Se o número máximo de iterações permitido é atingido, toma-se P^r como partição solução.

3.5 Algoritmo com ajustamento do número de classes: Fusão e Refinamento

Salvo conhecimento prévio do analista relativamente ao conjunto de dados, a definição *a priori* do número de classes da partição solução torna-se uma tarefa ingrata no processo de classificação não hierárquica, uma vez que a procura da "boa partição" teria que passar por sucessivas escolhas do valor de k . Na tentativa de ultrapassar este óbice, são introduzidos parâmetros e procedimentos no algoritmo básico, que permitem o ajuste do número de classes, tornando a procura da estrutura dos dados mais eficaz.

No algoritmo descrito na secção anterior, o número de classes é constante ao longo de todo o processo. Razões heurísticas justificam a introdução de procedimentos que permitem o controlo do número adequado de classes, dando assim origem à fusão e/ou refinamento de classes. Deste modo, dois parâmetros de controlo são usados no algoritmo: β e α , que representam os parâmetros de fusão e refinamento, respectivamente.

3.5.1 Fusão de classes

Após a classificação de todos os elementos, no final da etapa 1 do algoritmo, obtém-se uma partição em k classes, $P = \{A_1, \dots, A_k\}$. Para cada par de classes de P , A_j e $A_{j'}$, $j \neq j'$, de cardinais n_j e $n_{j'}$, respectivamente, calcula-se o valor da estatística em causa, U , sobre o conjunto $\{\gamma(I_i, I_{i'}) : (I_i, I_{i'}) \in A_j \times A_{j'}\}$, que tomamos como um conjunto de $n_j n_{j'}$ variáveis aleatórias independentes e identicamente distribuídas em $[0, 1]$, hipótese de referência adoptada nos métodos VL. O valor da respectiva função de distribuição de U é tomado como referência para avaliar a proximidade entre as classes. A fusão de A_j e $A_{j'}$ é levada a cabo se $\Gamma_U(A_j, A_{j'})$ é superior ao valor do parâmetro β .

Em geral, considerou-se $\beta = 0.95$. A decisão de fusão é tomada se $\Gamma_U(A_j, A_{j'}) > 0.95$. Está-se perante um teste de proximidade de classes com um nível de significância de 5%. Neste caso, obtém-se uma nova partição P' , em $k - 1$ classes:

$$P' = (P \setminus \{A_j, A_{j'}\}) \cup (A_j \cup A_{j'}).$$

3.5.2 Refinamento de classes

Sejam $I_i \in E$ e $\{A_1, \dots, A_k\}$ o conjunto de classes obtidas no momento anterior ao da classificação de I_i . Procedendo à classificação de I_i considera-se que o indivíduo se encontra demasiado afastado de todas as classes se

$$\max_{1 \leq j \leq k} \Gamma_U(\{I_i\}, A_j) < \alpha$$

onde U representa a estatística usada, formando-se uma nova classe constituída por um único elemento $\{I_i\}$. Assim, a partição obtida é dada por

$$P' = \{A_1, \dots, A_k, A_{k+1}\}, \text{ sendo } A_{k+1} = \{I_i\}.$$

Em geral, tomou-se $\alpha = 0.5$, isto é, se $\max_{1 \leq j \leq k} \Gamma_U(\{I_i\}, A_j) < 0.5$, o valor da estatística das semelhanças entre I_i e os elementos da classe A_j , tem maior probabilidade

de ser ultrapassada do que de não ser. Neste caso, toma-se I_i como um novo pólo de agregação.

Capítulo 4

Avaliação de Partições

Diferentes partições são obtidas quando sobre o mesmo conjunto de dados se aplicam diferentes métodos, ou diferentes decisões são tomadas ao longo de um procedimento classificatório. A definição de medidas de avaliação de uma partição facilita a procura da "melhor" partição no sentido do critério definido por essa medida.

Como já foi referido, a partição pretendida deverá satisfazer propriedades básicas de coesão interna e isolamento das classes. Nos algoritmos foram implementadas algumas medidas de avaliação da homogeneidade e isolamento de uma classe, que nas secções seguintes são expostas (cf. Gordon [17] e Hansen e Jaumard [19]).

4.1 Medidas de Homogeneidade de uma Classe

A primeira medida de homogeneidade apresentada, H_1 , destina-se apenas a conjuntos de dados descritos por variáveis quantitativas. As restantes baseiam-se numa matriz de proximidades.

Para o que se segue, considere-se A_j uma classe de cardinal n_j .

- **H_1 : Quadrado das Distâncias**

Esta medida de homogeneidade de uma classe é dada por

$$H_1(A_j) = \frac{1}{2n_j} \sum_{I_i, I_\ell \in A_j} d^2(I_i, I_\ell)$$

tratando-se d^2 do quadrado da distância euclideana. Uma classe A_j é tanto mais homogênea quanto menor o valor de H_1 .

- **H_2 : Medida Diâmetro**

A medida de homogeneidade "Diâmetro" baseia-se na dissemelhança entre os dois elementos mais afastados na classe. Logo, perante um quadro de dissemelhanças, a homogeneidade de A_j surge de

$$H_2(A_j) = \max \{ \gamma(I_i, I_\ell) : I_i, I_\ell \in A_j \}. \quad (4.1.1)$$

Neste caso, uma classe é mais homogênea do que outra se a medida diâmetro da primeira for inferior à medida diâmetro da segunda.

Se γ é um índice do tipo semelhança, (4.1.1) assume a seguinte forma:

$$H_2(A_j) = \min \{ \gamma(I_i, I_\ell) : I_i, I_\ell \in A_j \}. \quad (4.1.2)$$

- **H_3 : Medida Estrela**

Fixado um elemento da classe A_j , calcula-se a média das dissemelhanças aos restantes elementos da classe. A medida "Estrela" é dada pelo valor que minimiza a média destes valores, isto é,

$$H_3(A_j) = \min \left\{ \frac{1}{n_j - 1} \sum_{I_\ell \in A_j \setminus \{I_i\}} \gamma(I_i, I_\ell) : I_i \in A_j \right\}.$$

Uma classe será tanto mais homogênea quanto menor o valor da medida "Estrela". Tratando-se γ de uma medida de comparação do tipo semelhança, a medida "Estrela" é

$$H_3(A_j) = \max \left\{ \frac{1}{n_j - 1} \sum_{I_\ell \in A_j \setminus \{I_i\}} \gamma(I_i, I_\ell) : I_i \in A_j \right\}.$$

- H_4 : Soma das distâncias

Esta medida avalia a homogeneidade de uma classe através da média dos valores das dissemelhanças da matriz triangular inferior. Assim,

$$H_4(A_j) = 2 \sum_{I_i, I_\ell \in A_j, i > \ell} \frac{\gamma(I_i, I_\ell)}{n_j(n_j - 1)}.$$

A classe A_j será tanto mais homogênea quanto menor o valor de $H_4(A_j)$ (maior se γ for do tipo semelhança).

Os critérios desenvolvidos neste trabalho têm por base o uso de diferentes estatísticas de tendência central. Como foi já sugerido para a escolha dos centros iniciais, propõem-se adaptações das medidas de homogeneidade H_3 e H_4 dando origem a diferentes medidas que têm apenas significado nos respectivos critérios:

- Nos critérios NHVC e NHDC, sugere-se a seguinte adaptação na medida "Estrela": a homogeneidade de uma classe é dada pelo mínimo do centro amostral do conjunto de valores que medem a dissemelhança entre um elemento da classe e os restantes elementos da mesma classe, isto é,

$$H_5(A_j) = \min \left\{ \frac{\max_{I_\ell \in A_j \setminus \{I_i\}} \gamma(I_i, I_\ell) + \min_{I_\ell \in A_j \setminus \{I_i\}} \gamma(I_i, I_\ell)}{2} : I_i \in A_j \right\}.$$

No caso da medida H_4 , a alteração proposta é dada pelo centro amostral dos valores das dissemelhanças entre elementos que pertencem à mesma classe, da matriz triangular inferior

$$H_6(A_j) = \frac{\max \gamma(I_i, I_\ell) + \min \gamma(I_i, I_\ell)}{2}, \quad I_i, I_\ell \in A_j, \quad i > \ell.$$

- Nos critérios NHVMED e NHD MED, sugerem-se adaptações às medidas H_3 e H_4 , dando assim origem às medidas H_7 e H_8 . Estas assumem a seguinte forma:

$$H_7(A_j) = \min \{ \text{mediana} \{ \gamma(I_i, I_\ell) : I_\ell \in A_j \setminus \{I_i\} \} : I_i \in A_j \}.$$

e

$$H_8(A_j) = \text{mediana} \{ \gamma(I_i, I_\ell) : I_i, I_\ell \in A_j, i > \ell \}.$$

Assim, H_7 é o valor mínimo da mediana das dissemelhanças entre um elemento e os restantes elementos da mesma classe. A medida de homogeneidade H_8 é dada pela mediana das dissemelhanças na matriz triangular inferior entre elementos da mesma classe.

- Nos critérios NIIVGM e NIIDGM, analogamente ao que já foi apresentado, são propostas duas novas medidas baseadas em H_3 e H_4 , que têm por base o uso da estatística média geométrica. Desta forma, H_9 é dada pelo mínimo da média dos logaritmos dos valores das dissemelhanças entre elementos da classe. Considerando os valores simétricos do logaritmo das dissemelhanças entre um dado elemento da classe e os restantes elementos da mesma classe, H_9 é dada pelo máximo da referida média.

Analogamente, a medida H_{10} é dada pela média dos logaritmos dos valores de comparação entre elementos da mesma classe da matriz triangular inferior. Assim,

$$H_9(A_j) = \max \left\{ \frac{\sum_{I_\ell \in A_j \setminus \{I_i\}} -\ln \gamma(I_i, I_\ell)}{n_j - 1} : I_i \in A_j \right\}$$

e

$$H_{10}(A_j) = \frac{2 \sum_{I_i, I_\ell \in A_j, i > \ell} -\ln \gamma(I_i, I_\ell)}{n_j(n_j - 1)}.$$

No método NIIMEAN mantêm-se os critérios H_3 e H_4 , dado que ambos os critérios se baseiam na média aritmética dos valores do índice de comparação entre pares de elementos.

4.2 Medidas de Isolamento

De seguida, apresentam-se duas medidas de isolamento, Iso .

A primeira medida de isolamento de uma classe, Iso_1 , é dada pela menor dissemelhança entre um indivíduo da classe e um indivíduo exterior à classe, isto é,

$$Iso_1(A_j) = \min \{ \gamma(I_i, I_\ell) : I_i \in A_j, I_\ell \notin A_j \}.$$

Tratando-se γ do tipo semelhança, a medida de isolamento da classe é dada por

$$Iso_1(A_j) = \max \{ \gamma(I_i, I_\ell) : I_i \in A_j, I_\ell \notin A_j \}. \quad (4.2.1)$$

A segunda medida introduzida no algoritmo mede o isolamento de uma classe a partir da média de dissemelhanças (ou semelhanças) entre indivíduos da classe e indivíduos não pertencentes à mesma classe. Concretizando,

$$Iso_2(A_j) = \sum_{I_i \in A_j} \sum_{I_\ell \notin A_j} \frac{\gamma(I_i, I_\ell)}{n_j(n - n_j)}. \quad (4.2.2)$$

Sendo γ do tipo dissemelhança, baixos valores de Iso_1 e Iso_2 indicam classes pouco isoladas.

Analogamente ao que feito para as medidas de homogeneidade também estas duas medidas de isolamento poderiam facilmente ser adaptadas aos respectivos métodos.

Para uma exposição mais detalhada sobre estas e outras medidas de homogeneidade e isolamento de uma classe, consulte-se Hansen e Jaumard [19].

As medidas de homogeneidade e isolamento apresentadas permitem avaliar classes de uma partição. A combinação dos valores destas medidas dará facilmente origem a medidas que permitem avaliar a partição na sua globalidade, e que na secção seguinte são referidas.

4.3 Avaliação Global de uma Partição

Um dos critérios de avaliação da qualidade da partição, implementado no algoritmo, consiste na soma da medida de homogeneidade de todas as classes da partição, isto é,

$$G_1 = \sum_{j=1}^k H(A_j).$$

Outra medida da qualidade da partição igualmente usada, é dada pelo valor mínimo da homogeneidade das classes da partição.

$$G_2 = \min \{ H(A_j) : j = 1, \dots, k \}.$$

No algoritmo implementado para o método NHMEAN por Brito [9] um critério usado para avaliar a qualidade da partição, segundo a filosofia da metodologia Validade de Ligação, consiste no valor da função de distribuição da média das semelhanças VL entre os pares de elementos pertencentes à mesma classe. Isto é,

$$G_3 = F \left(\frac{\sum_{(I_i, I_\ell) \in E \times E} \gamma(I_i, I_\ell) Ind_{I_i, I_\ell}}{\sum_{j=1}^k \frac{n_j(n_j - 1)}{2}} \right) \quad (4.3.1)$$

sendo F a função de distribuição da estatística média aritmética amostral, n_j o cardinal da j -ésima classe da partição e Ind_{I_i, I_ℓ} a função indicatriz da partição, isto é,

$$Ind_{I_i, I_\ell} = \begin{cases} 1, & \text{se } I_i \text{ e } I_\ell \text{ pertencem à mesma classe} \\ 0, & \text{c. c.} \end{cases}$$

Na hipótese de referência atrás referida, F é dada por (3.1.2).

A qualidade da partição é tanto melhor, quanto maior o valor indicado em (4.3.1).

A expressão em (4.3.1) tem por base a estatística média aritmética amostral. Mais uma vez, razões heurísticas levaram a adaptações deste critério de acordo com as estatísticas usadas nos métodos abordados. Os critérios que de seguida se apresentam dão continuidade à linha de pensamento da metodologia VL.

No método NIIVC, considerou-se como critério global de qualidade da partição o valor da função de distribuição do centro das semelhanças entre pares de elementos da mesma classe, isto é,

$$G_4 = F \left(\frac{\min \gamma(I_i, I_\ell) + \max \gamma(I_i, I_\ell)}{2}, (I_i, I_\ell) \in A_j \times A_j, i < \ell, 1 \leq j \leq k \right) \quad (4.3.2)$$

sendo F a função dada em (3.1.14).

Para o método NIIVMED, a partição é avaliada na sua globalidade por

$$G_5 = F(\text{mediana}\{\gamma(I_i, I_\ell) : (I_i, I_\ell) \in A_j \times A_j, i < \ell, 1 \leq j \leq k\}).$$

Neste caso, a expressão de F depende da dimensão do conjunto de pares de elementos pertencentes à mesma classe. Se o cardinal de $\{\gamma(I_i, I_\ell) : (I_i, I_\ell) \in A_j \times A_j, i < \ell, 1 \leq j \leq k\}$ é par então F é a função expressa em (3.1.13), se é ímpar F é a função em (3.1.12).

Por último, no método NHVGM, a avaliação da partição obtém-se do valor da função de distribuição da média geométrica das semelhanças VL entre todos os pares de elementos da mesma classe. Considerando o simétrico dos valores da função de comparação entre elementos nos seus logaritmos, obtém-se

$$G_6 = F \left(\frac{\sum_{(I_i, I_\ell) \in E \times E} -\text{Ind}_{I_i, I_\ell} \ln \gamma(I_i, I_\ell)}{\sum_{j=1}^k \frac{n_j(n_j - 1)}{2}} \right)$$

onde F é a função representada em (3.1.8).

No caso do conjunto a avaliar ser razoavelmente grande, o número de comparações entre elementos da mesma classe da partição atinge um valor elevado, pelo que para valores de abcissas superiores a 0.5, o valor da função de distribuição nos métodos NHMEAN, NHVGM e NHVC, assume valores muito próximos de 1; quando o valor da estatística amostral é inferior a 0.5 a função de distribuição assume valores próximos de 0, fragilizando a qualidade da medida de avaliação global da partição.

4.4 Comparação de Partições

Como já se fez notar, em métodos de Classificação Não Hierárquica, tomadas de decisão, tais como o número de classes pretendido, o processo de inicialização, o critério a aplicar, ou mesmo, os valores dos parâmetros de convergência, condicionam fortemente a partição final dando assim origem a diferentes partições. Estes factos tornam pertinentes a avaliação das escolhas realizadas! Neste sentido, a avaliação da escolha da partição solução que optimiza a proximidade com a "boa" partição (se conhecida), considerada como aquela que traduz mais fielmente a estrutura dos dados, poderá ser feita tendo em conta índices de semelhança entre partições. De facto, este assunto tem sido objecto de estudo por parte de alguns especialistas, tornando-se tema recorrente

na literatura mais recente sobre Classificação.

A medida de comparação de partições usada neste trabalho foi o **Índice de Rand Modificado** (Hubert e Arabie [21]).

Considerem-se P_1 e P_2 , duas partições de E , em k_1 e k_2 classes, respectivamente. Seja $n_{jj'}$ o número de elementos comuns às classes A_j^1 e $A_{j'}^2$, com $j = 1, \dots, k_1$, $j' = 1, \dots, k_2$, pertencentes às partições P_1 e P_2 , respectivamente. Os $\frac{n(n-1)}{2}$ pares não ordenados de elementos distintos de E subdividem-se em quatro tipos:

- i) Os pares que pertencem à mesma classe em P_1 e P_2 ;
- ii) Os pares que pertencem a diferentes classes em P_1 e P_2 ;
- iii) Os pares que pertencem à mesma classe em P_1 e a diferentes classes em P_2 ;
- vi) Os pares que pertencem à mesma classe em P_2 e a diferentes classes em P_1 .

Seja A o número total de pares de elementos do primeiro e segundo tipo, isto é, o número de pares de elementos com igual tratamento (acordos), e D o número de elementos do terceiro e quarto tipo, ou seja, o número de pares com diferente tratamento (desacordos). Intuitivamente, tem-se que se duas partições são semelhantes então obtêm-se grandes valores de A e pequenos valores de D .

Dada a tabela de contingência 4.1 com $n_{j.} = \sum_{j'=1}^{k_2} n_{jj'}$ e $n_{.j'} = \sum_{j=1}^{k_1} n_{jj'}$, tem-se que

classes	A_1^2	$\dots$	$A_{k_2}^2$
A_1^1	n_{11}	$\dots$	n_{1k_2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$A_{k_1}^1$	$n_{k_1 1}$	$\dots$	$n_{k_1 k_2}$

Tabela 4.1: Tabela de Contingência

(cf. Hubert e Arabie [21]):

$$A = \binom{n}{2} + 2 \sum_{j=1}^{k_1} \sum_{j'=1}^{k_2} \binom{n_{jj'}}{2} - \left[\sum_{j=1}^{k_1} \binom{n_{j.}}{2} + \sum_{j'=1}^{k_2} \binom{n_{.j'}}{2} \right]. \quad (4.4.1)$$

Rand (cf. Hubert e Arabie [21]) propõe como medida de semelhança entre duas partições a percentagem de pares com igual tratamento, isto é,

$$R = A / \binom{n}{2} = \frac{2A}{n(n-1)}. \quad (4.4.2)$$

Se R tomar valores próximos de 1, tem-se um elevado grau de semelhança entre as duas partições. Se R assumir o valor 0, as partições não apresentam qualquer semelhança. Note-se que, tal como é definido, o Índice de Rand atribui a mesma importância aos pares de elementos que pertencem à mesma classe em P_1 e P_2 e aos pares de elementos que estão em diferentes classes em P_1 e P_2 .

Para além da medida de Rand, baseada na probabilidade da ocorrência de pares de elementos com igual tratamento, outras medidas de comparação foram já usadas: $D / \binom{n}{2}$ que é a medida de probabilidade de ocorrência de pares de elementos com diferente tratamento e $(A - D) / \binom{n}{2}$, diferença entre a probabilidade de acordos e a probabilidade de desacordos.

Seja $W = \sum_{j=1}^{k_1} \sum_{j'=1}^{k_2} \binom{N_{jj'}}{2}$ a variável aleatória que representa o número de pares de elementos do tipo i), isto é, pares de elementos que se encontram na mesma classe nas duas partições. Sob a hipótese da escolha aleatória de duas partições do mesmo conjunto seguir uma distribuição hipergeométrica tem-se que (cf. Hubert e Arabie [21])

$$E(W) = \sum_{j=1}^{k_1} \binom{n_{j\cdot}}{2} \sum_{j'=1}^{k_2} \binom{n_{\cdot j'}}{2} / \binom{n}{2}. \quad (4.4.3)$$

Por (4.4.1) e (4.4.3) o valor esperado do Índice de Rand é

$$\begin{aligned} E\left(A / \binom{n}{2}\right) &= 1 + 2 \sum_{j=1}^{k_1} \binom{n_{j\cdot}}{2} \sum_{j'=1}^{k_2} \binom{n_{\cdot j'}}{2} / \binom{n}{2}^2 \\ &\quad - \left[\sum_{j=1}^{k_1} \binom{n_{j\cdot}}{2} + \sum_{j'=1}^{k_2} \binom{n_{\cdot j'}}{2} \right] / \binom{n}{2}. \end{aligned}$$

Usando a correcção estatística

$$\frac{Indice - E(Indice)}{Max_{Indice} - E(Indice)} \quad (4.4.4)$$

em que $E(Indice)$ é o valor esperado do índice e Max_{Indice} é o valor máximo que este toma, Hubert e Arabie [21] propõem uma correcção sobre a medida expressa em (4.4.2), dando assim origem ao Índice de Rand Modificado, IRM , que mede a semelhança entre partições pela expressão:

$$IRM(P_1, P_2) = \frac{\sum_{j=1}^{k_1} \sum_{j'=1}^{k_2} \binom{n_{jj'}}{2} - \sum_{j=1}^{k_1} \binom{n_{j.}}{2} \sum_{j'=1}^{k_2} \binom{n_{.j'}}{2} / \binom{n}{2}}{\left[\sum_{j=1}^{k_1} \binom{n_{j.}}{2} + \sum_{j'=1}^{k_2} \binom{n_{.j'}}{2} \right] / 2 - \sum_{j=1}^{k_1} \binom{n_{j.}}{2} \sum_{j'=1}^{k_2} \binom{n_{.j'}}{2} / \binom{n}{2}}.$$

Tem-se que IRM toma o valor máximo 1 e $IRM = 0$ quando o valor do índice é igual ao valor esperado. No caso das partições serem muito diferentes o valor de IRM pode tomar valores negativos.

A correcção estatística proposta por Hubert e Arabie vem, segundo estes, aumentar a sensibilidade do índice perante diferenças significativas entre as duas partições em comparação.

Outros índices de comparação de partições foram estudados num recente texto de Saporta e Youness [38]. Este trabalho engloba o estudo do comportamento dos dois índices que aqui são usados, Índice de Rand e Índice de Rand Modificado, e também a versão assimétrica do Índice de Rand, apropriado para medir a concordância entre duas partições com um diferente número de classes. O texto avalia ainda o desempenho de dois novos índices: um baseado no teste de Mac Nemar, e o outro no Índice de Jacard. O primeiro é um teste não paramétrico que pretende testar contagens ou proporções em duas amostras emparelhadas. O segundo índice surge da analogia com o já referido Índice de Jacard para dados binários. Segue-se uma breve exposição do Índice de Rand Assimétrico, do Índice de MacNemar e do Índice de Jacard.

Índice de Rand Assimétrico

Como já foi referido, o Índice de Rand Assimétrico, *IRA*, usa-se no caso de se estar perante duas partições de um mesmo conjunto de indivíduos, com um número de classes diferente.

Sejam P_1 e P_2 duas partições de um mesmo conjunto de n indivíduos, em k_1 e k_2 classes, respectivamente, com $k_1 > k_2$. Diz-se que P_1 é uma partição mais fina que P_2 se dois quaisquer indivíduos da mesma classe em P_1 , estão também na mesma classe em P_2 , isto é, $\forall j = 1, \dots, k_1, \exists j' : A_j^1 \subseteq A_{j'}^2$, onde A_j^1 designa a j -ésima classe de P_1 e $A_{j'}^2$ designa a j' -ésima classe de P_2 .

O Índice de Rand na sua versão assimétrica pretende avaliar até que ponto uma partição P_1 está incluída na partição P_2 , e é dado por:

$$IRA(P_1, P_2) = 1 + \frac{\sum_{j=1}^{k_1} \sum_{j'=1}^{k_2} \binom{n_{jj'}}{2} - \sum_{j=1}^{k_1} \binom{n_j}{2}}{\binom{n}{2}}$$

sendo $n_{jj'}$ o número de elementos que estão simultaneamente na j -ésima classe de P_1 e na j' -ésima classe de P_2 .

Tem-se que $0 \leq IRA \leq 1$. Se P_1 é mais fina do que P_2 , então $IRA = 1$.

Considerando os n^2 pares de indivíduos (I_i, I_ℓ) , onde se distinguem os pares (I_i, I_ℓ) e (I_ℓ, I_i) e se contabilizam os pares (I_i, I_i) , obtém-se (cf. Saporta e Youness [38]):

$$IRA(P_1, P_2) = \frac{n^2 + \sum_{j=1}^{k_1} \sum_{j'=1}^{k_2} n_{jj'}^2 - \sum_{j=1}^{k_1} n_j^2}{n^2}.$$

Note-se que no caso de P_1 e P_2 terem o mesmo número de classes, o Índice de Rand Assimétrico não coincide com o Índice de Rand.

Usando a correcção estatística em (4.4.4), surge o Índice de Rand Assimétrico Modificado, *IRAM*, dado por (cf. Saporta e Youness [38])

$$IRAM(P_1, P_2) = \frac{n^2 \sum_{j=1}^{k_1} \sum_{j'=1}^{k_2} n_{jj'}^2 - \sum_{j=1}^{k_1} n_{j.}^2 \sum_{j'=1}^{k_2} n_{.j'}}{n^2 \sum_{j=1}^{k_1} n_{j.}^2 - \sum_{j=1}^{k_1} n_{j.}^2 \sum_{j'=1}^{k_2} n_{.j'}}.$$

Índice de Mac Nemar

O teste de Mac Nemar é um teste não paramétrico utilizado para verificar a igualdade de proporções em amostras emparelhadas para variáveis qualitativas ordinais. Este teste é particularmente apropriado para testar a significância de mudanças de estado ou opinião. Por exemplo, considere-se a opinião de uma amostra de n indivíduos relativamente a um produto antes e após uma campanha publicitária. Dada a tabela de contingência em 4.2 onde o estado "1" representa opinião favorável, o estado "0" repre-

$\begin{array}{c c} D \\ \hline A \end{array}$	1	0
1	a	b
0	c	d

Tabela 4.2: Tabela de Contingência

senta opinião desfavorável, a (respectivamente, d) representa o número de indivíduos que mantêm a mesma opinião favorável (respectivamente, desfavorável) antes e após a campanha publicitária, b e c representam o número de indivíduos que mudaram de opinião, pretende-se testar a hipótese nula de a probabilidade de mudança do estado "0" para "1", ser igual à probabilidade de mudança do estado "1" para "0". A estatística de teste usada é

$$M_c = \frac{c - b}{\sqrt{c + b}}. \quad (4.4.5)$$

Adaptando o Índice de Mac Nemar, IM_c , ao conjunto de pares de elementos, obtém-se um novo modo de medir a concordância entre as duas partições, que consiste em averiguar se os pares que são separados, o são por "acaso". Isto é, estuda-se o desacordo entre os pares de elementos.

Quando o valor do IM_c se aproxima de 0, conclui-se que a diferença entre as partições é devida ao acaso. Atendendo à expressão em (4.4.5) e considerando os n^2 pares de indivíduos, tem-se

$$\begin{aligned}
 IM_c(P_1, P_2) &= \frac{\sum_{j=1}^{k_1} n_{j\cdot}^2 - \sum_{j'=1}^{k_2} n_{\cdot j'}^2}{2 \sqrt{\frac{1}{2} \left(\sum_{j=1}^{k_1} n_{j\cdot}^2 + \sum_{j'=1}^{k_2} n_{\cdot j'}^2 \right) - \sum_{j=1}^{k_1} \sum_{j'=1}^{k_2} n_{jj'}^2}} \\
 &= \frac{\sum_{j=1}^{k_1} n_{j\cdot}^2 - \sum_{j'=1}^{k_2} n_{\cdot j'}^2}{\sqrt{2} \sqrt{\sum_{j=1}^{k_1} n_{j\cdot}^2 + \sum_{j'=1}^{k_2} n_{\cdot j'}^2 - 2 \sum_{j=1}^{k_1} \sum_{j'=1}^{k_2} n_{jj'}^2}}.
 \end{aligned}$$

Índice de Jacard

O Índice de Jacard é um índice de concordância usado para medir a semelhança entre pares de elementos descritos por variáveis binárias. Perante uma tabela de contingência como a que figura na tabela 1.1, este índice de semelhança é dado por $s(I_i, I_\ell) = \frac{a}{a + b + c}$.

Dadas duas partições P_1 e P_2 de um mesmo conjunto de n indivíduos, o Índice de Jacard, IJ , usado para medir a semelhança entre P_1 e P_2 , é dado por

$$IJ(P_1, P_2) = \frac{\sum_{j=1}^{k_1} \sum_{j'=1}^{k_2} n_{jj'}^2 - n}{\sum_{j=1}^{k_1} n_{j\cdot}^2 + \sum_{j'=1}^{k_2} n_{\cdot j'}^2 - \sum_{j=1}^{k_1} \sum_{j'=1}^{k_2} n_{jj'}^2 - n}.$$

O desempenho dos índices expostos foi estudado em conjuntos de dados obtidos de acordo com um modelo de k classes latentes descritas por p variáveis numéricas de distribuição normal (Saporta e Youness [38]). Assim, agrupam-se as p variáveis em dois grupos e classifica-se o mesmo conjunto de n indivíduos de acordo com estes dois grupos de variáveis, usando o método das k -médias de MacQueen.

O trabalho de Saporta e Youness [38] permitiu concluir que a distribuição do Índice de Rand depende da separação das classes, do número de indivíduos e do número de classes da partição. Alguns exemplos formados por classes com diferentes graus de separação mostraram que o Índice de Rand depende mais fortemente do número de indivíduos e de classes do que do grau de separação entre classes. Relativamente à distribuição do Índice de Mac Nemar obtida por simulação, os resultados mostraram que a maioria dos valores obtidos nas 1000 réplicas, se encontram muito próximo de zero, não se rejeitando a hipótese nula. Já o Índice de Jacard sugere uma distribuição bimodal para o exemplo analisado. Foi também realizado um estudo da média do Índice de Rand, do Índice de Mac Nemar e do Índice de Jacard, variando o número de classes latentes equiprováveis, o que permitiu concluir que o Índice de Mac Nemar e o Índice de Rand variam no mesmo sentido do número de classes, ao contrário do Índice de Jacard que varia em sentido inverso.

Em Guénoche e Garreta [18] são propostas novas medidas de avaliação de classes e de comparação de partições, desenvolvidas no contexto da teoria de grafos.

Capítulo 5

Aplicações

O estudo do desempenho dos métodos propostos contempla a aplicação destes métodos quer a dados reais, quer a dados simulados. Neste capítulo apresentam-se os resultados obtidos nas experiências levadas a cabo com os novos métodos propostos e também com o método NHMEAN e o método das k -médias de MacQueen.

Para os diferentes critérios expostos, compara-se a partição solução com a partição de referência, recorrendo ao Índice de Rand Modificado. O conhecimento prévio da estrutura dos dados fornece a partição de referência.

5.1 Dados Simulados

O desempenho dos métodos explorados é analisado face a estruturas classificatórias simuladas. A aplicação dos métodos em causa é efectuada a conjuntos de dados gerados em $\mathbb{R}^2$, com distribuição binormal. Nos exemplos tratados, os conjuntos são formados por três classes de igual dimensão, com 50 elementos cada. Analisam-se conjuntos de classes com formas distintas, duas classes elípticas e uma classe esférica, e com diferentes graus de separação.

Descreve-se sucintamente o método usado na geração de variáveis aleatórias multinormais, procedendo-se posteriormente à discussão dos resultados.

5.1.1 Distribuição Multinormal

Seja $X = (X_1, \dots, X_p)$ um vector aleatório que segue uma distribuição multinormal de média $\mu = (\mu_1, \dots, \mu_p)$ e matriz de covariâncias $\Sigma =$

$$\begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{bmatrix},$$

$X \sim N(\mu, \Sigma)$. Sendo Σ uma matriz definida positiva e simétrica, existe uma matriz triangular inferior $T =$

$$\begin{bmatrix} t_{11} & 0 & \cdots & 0 \\ t_{21} & t_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ t_{p1} & t_{p2} & \cdots & t_{pp} \end{bmatrix}$$

tal que, $\Sigma = TT^t$.

O vector aleatório X pode ser obtido da igualdade

$$X = TZ + \mu$$

sendo $Z = (Z_1, \dots, Z_p)$ um vector com distribuição multinormal de média nula, e cuja matriz de covariâncias é a matriz identidade $I_{p \times p}$, ou seja, Z_j , com $j = 1, \dots, p$, segue uma distribuição normal estandardizada, $Z_j \sim N(0, 1)$. Os elementos da matriz T obtêm-se da fórmula de recorrência¹

$$t_{jj'} = \frac{\sigma_{jj'} - \sum_{m=1}^{j'-1} t_{jm} t_{j'm}}{\sqrt{\sigma_{j'}^2 - \sum_{m=1}^{j'-1} t_{j'm}^2}}$$

para $2 \leq j' \leq j \leq p$. Para $j' = 1$ tem-se $t_{j1} = \frac{\sigma_{j1}}{\sqrt{\sigma_1^2}}$, com $1 \leq j \leq p$.

Para a geração da distribuição normal estandardizada é usado o método sugerido por Box e Müller (cf. Rubinstein [37]): sendo U_1 e U_2 , duas variáveis aleatórias uniformes em $[0, 1]$, então Z_1 e Z_2 definidas por

$$Z_1 = \sqrt{-2 \ln U_1} \cos 2\pi U_2 \text{ e } Z_2 = \sqrt{-2 \ln U_1} \sin 2\pi U_2$$

são variáveis aleatórias com distribuição normal estandardizada.

¹Considera-se $\sigma_{jj} = \sigma_j^2$ para $j = 1, \dots, p$.

5.1.2 Resultados

O critério de escolha dos centros iniciais foi o critério *wise-choice* com respectivas adaptações, sendo $L = 5$, a dimensão da L -vizinhança de cada indivíduo. O índice de semelhança usado foi $s(I_i, I_\ell) = \frac{1 + \cos \alpha}{2}$, sendo α o ângulo formado por I_i e I_ℓ .

A comparação das partições obtidas com a partição gerada é feita com base no Índice de Rand Modificado, apresentando-se a média e desvio-padrão dos valores deste índice para 100 réplicas.

I. Exemplo 1

Considere-se uma amostra com distribuição binormal obtida de acordo com os seguintes parâmetros:

Classe A : $\mu_1 = 0$; $\mu_2 = 0$; $\sigma_1^2 = 0.7$; $\sigma_{12} = 1$; $\sigma_2^2 = 1.5$;

Classe B : $\mu_1 = 3$; $\mu_2 = 6$; $\sigma_1^2 = 0.5$; $\sigma_{12} = 0$; $\sigma_2^2 = 0.5$;

Classe C : $\mu_1 = -5$; $\mu_2 = 3$; $\sigma_1^2 = 0.9$; $\sigma_{12} = -0.7$; $\sigma_2^2 = 0.8$;

representada na figura 5.1.

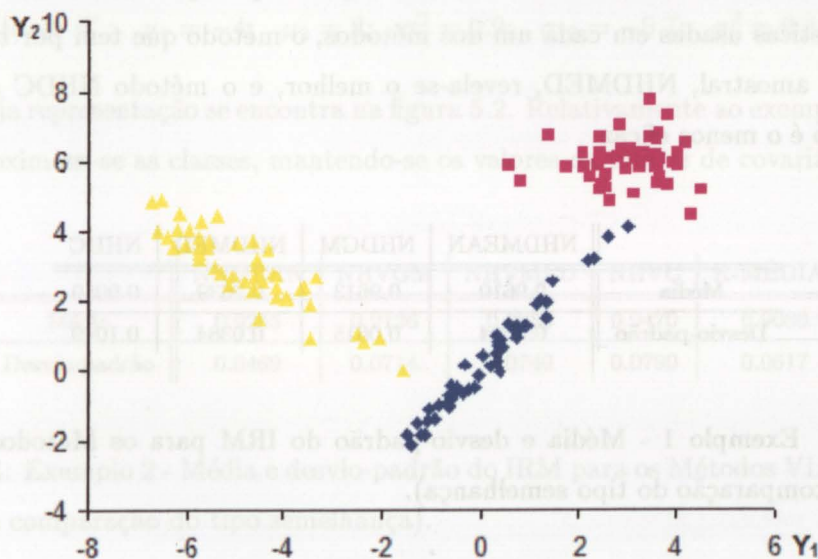


Figura 5.1: Conjunto de dados do Exemplo 1.

Na tabela 5.1 verifica-se que o valor médio do Índice de Rand Modificado é bastante satisfatório para qualquer dos métodos. O método NHMEAN é aquele que tem melhor desempenho na classificação deste conjunto, seguido de perto do

	NHMEAN	NHVGM	NHVMED	NHVC	K-MÉDIAS
Média	0.9679	0.9445	0.9556	0.9580	0.9674
Desvio-padrão	0.0312	0.0643	0.0348	0.0970	0.0325

Tabela 5.1: Exemplo 1 - Média e desvio-padrão do IRM para os Métodos VL e k -Médias (Índice de comparação do tipo semelhança).

método das k -médias. De entre os novos métodos VL propostos neste trabalho, o método NHVC, foi o método que se mostrou ligeiramente mais eficaz. O elevado desvio-padrão obtido no método NHVC justificou uma análise mais cuidada sobre o valor do Índice de Rand Modificado obtido em cada uma das 100 réplicas: em algumas das experiências realizadas, verificou-se que uma má escolha de centros iniciais levou ao isolamento destes pólos de agregação em classes singulares, tendo como consequência a diminuição do valor do IRM.

A tabela 5.2 mostra o desempenho dos métodos directos sobre a mesma nuvem de pontos. Tendo como critério de afectação a aplicação directa das diferentes estatísticas usadas em cada um dos métodos, o método que tem por base a mediana amostral, NHDMED, revela-se o melhor, e o método NHDC que usa o centro é o menos eficaz.

	NHDMEAN	NHDGM	NHDMED	NHDC
Média	0.9610	0.9613	0.9639	0.9050
Desvio-padrão	0.0604	0.0615	0.0384	0.1049

Tabela 5.2: Exemplo 1 - Média e desvio-padrão do IRM para os Métodos Directos (Índice de comparação do tipo semelhança).

Os métodos directos foram ainda testados usando como medida de comparação entre elementos a distância euclideana. Analisando a tabela 5.3 verifica-se que se mantém o melhor desempenho do método NHDMED seguido do método NHDGM.

Comparando os resultados dos métodos VL com os dos métodos directos, verificou-se que no exemplo em estudo, o uso da função de distribuição é mais

	NHDMEAN	NHDGM	NHDMED	NHDC
Média	0.9623	0.9638	0.9745	0.8868
Desvio-padrão	0.0563	0.0667	0.040	0.1099

Tabela 5.3: Exemplo 1 - Média e desvio-padrão do IRM para os Métodos Directos (Índice de comparação do tipo dissimilaridade).

vantajosa no caso do método ter por base a estatística média aritmética ou o centro amostral, por outro lado, o uso directo da estatística mediana ou média geométrica fornece resultados ligeiramente superiores aos resultados dos respectivos métodos VL.

II. Exemplo 2

Neste exemplo é estudado o desempenho dos métodos VL num conjunto de pontos que distingue três classes geradas de acordo com os seguintes parâmetros:

Classe A : $\mu_1 = 0; \mu_2 = 0; \sigma_1^2 = 0.7; \sigma_{12} = 1; \sigma_2^2 = 1.5;$
Classe B : $\mu_1 = 2; \mu_2 = 5; \sigma_1^2 = 0.5; \sigma_{12} = 0; \sigma_2^2 = 0.5;$
Classe C : $\mu_1 = -4; \mu_2 = 2; \sigma_1^2 = 0.9; \sigma_{12} = -0.7; \sigma_2^2 = 0.8,$

e cuja representação se encontra na figura 5.2. Relativamente ao exemplo anterior, aproximam-se as classes, mantendo-se os valores da matriz de covariâncias.

	NHMEAN	NHVGGM	NHVMED	NHVC	K-MÉDIAS
Média	0.9255	0.9126	0.8859	0.9470	0.9086
Desvio-padrão	0.0469	0.0714	0.0740	0.0790	0.0617

Tabela 5.4: Exemplo 2 - Média e desvio-padrão do IRM para os Métodos VL e *k*-Médias (Índice de comparação do tipo similaridade).

Os valores que figuram na tabela 5.4 continuam a ser satisfatórios. Neste caso, o método NHVC é o que melhores resultados fornece ainda que apresente maior desvio-padrão.

A aplicação dos métodos directos conduziu aos resultados expressos na tabela 5.5.

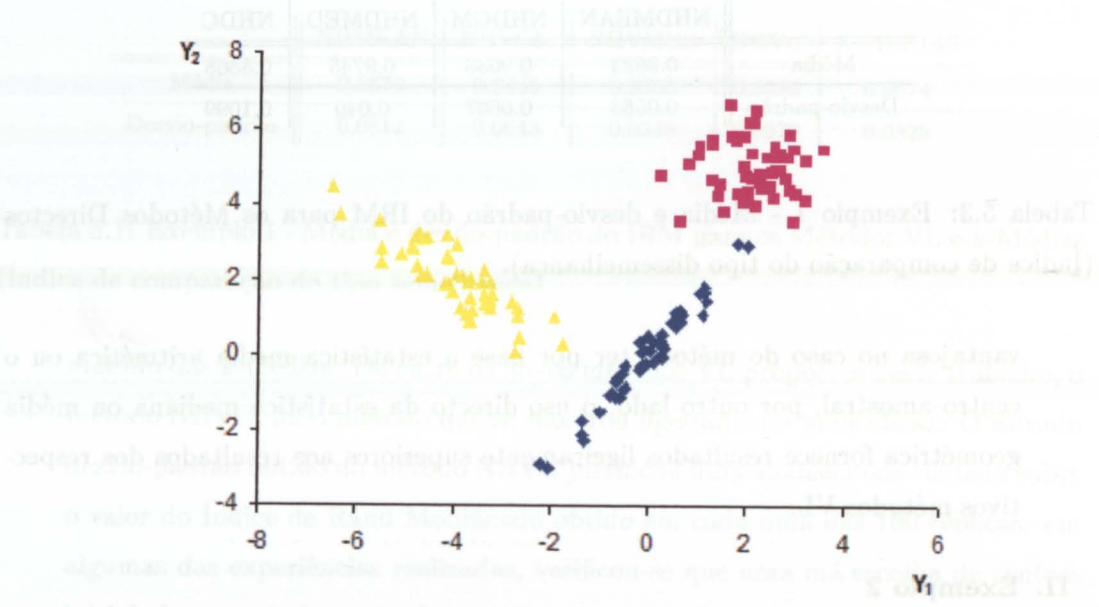


Figura 5.2: Conjunto de dados do Exemplo 2.

	NHDMEAN	NHDGM	NHDMED	NHDC
Média	0.8345	0.7192	0.8781	0.7969
Desvio-padrão	0.1379	0.1430	0.0798	0.1037

Tabela 5.5: Exemplo 2 - Média e desvio-padrão do IRM para os Métodos Directos (Índice de comparação do tipo semelhança).

Comparando os valores do IRM da tabela 5.5 com os valores do IRM da tabela 5.4, conclui-se que no exemplo em estudo é mais vantajosa a aplicação dos métodos VL face à aplicação dos respectivos métodos directos.

III. Exemplo 3

Outro exemplo tratado considera o conjunto representado na figura 5.3, cujas classes são obtidas segundo os seguintes parâmetros:

- Classe A : $\mu_1 = 0;$ $\mu_2 = 0;$ $\sigma_1^2 = 0.3;$ $\sigma_{12} = 0.3;$ $\sigma_2^2 = 0.5;$
- Classe B : $\mu_1 = 2;$ $\mu_2 = 5;$ $\sigma_1^2 = 0.25;$ $\sigma_{12} = 0;$ $\sigma_2^2 = 0.25;$
- Classe C : $\mu_1 = -4;$ $\mu_2 = 2;$ $\sigma_1^2 = 0.6;$ $\sigma_{12} = -0.4;$ $\sigma_2^2 = 0.3.$

Comparativamente ao exemplo anterior, o conjunto aqui estudado é formado por classes com o mesmo grau de separação e menos alongadas.

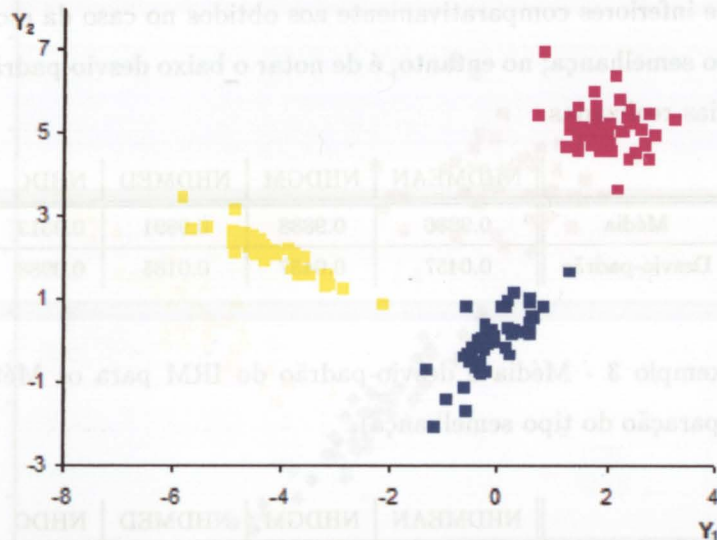


Figura 5.3: Conjunto de dados do Exemplo 3.

Os métodos apresentam em média, um valor do IRM superior a 0.94 (tabela 5.6). Os métodos *k*-médias e NHMEAN são os que se manifestam mais eficazes, ainda que os resultados obtidos com o método NHVMED estejam muito próximos dos bons resultados que estes dois produziram. O método NHVC apresenta novamente o maior desvio-padrão.

	NHMEAN	NHVGm	NHVMED	NHVC	K-MÉDIAS
Média	0.9739	0.9460	0.9710	0.9409	0.9903
Desvio-padrão	0.0248	0.0487	0.0260	0.1306	0.0189

Tabela 5.6: Exemplo 3 - Média e desvio-padrão do IRM para os Métodos VL e *k*-Médias (Índice de comparação do tipo semelhança).

Os resultados obtidos com a aplicação dos métodos directos constam da tabela 5.7. Estes são bastante satisfatórios realçando-se, contudo, o facto de o uso da função de distribuição do centro amostral produzir melhores resultados face à aplicação directa desta estatística sobre o conjunto das semelhanças VL.

O uso do índice de comparação de tipo VL que tem por base a escolha da distância euclideana conduziu aos resultados expressos na tabela 5.8, na aplicação

dos métodos directos. À excepção do método NHDC, os valores do IRM são sensivelmente inferiores comparativamente aos obtidos no caso da escolha do índice ser do tipo semelhança; no entanto, é de notar o baixo desvio-padrão do IRM nas experiências realizadas.

	NHDMEAN	NHDGM	NHDMED	NHDC
Média	0.9886	0.9888	0.9891	0.9313
Desvio-padrão	0.0457	0.0457	0.0185	0.0988

Tabela 5.7: Exemplo 3 - Média e desvio-padrão do IRM para os Métodos Directos (Índice de comparação do tipo semelhança).

	NHDMEAN	NHDGM	NHDMED	NHDC
Média	0.9884	0.9867	0.9895	0.9721
Desvio-padrão	0.0162	0.0182	0.0186	0.0364

Tabela 5.8: Exemplo 3 - Média e desvio-padrão do IRM para os Métodos Directos (Índice de comparação do tipo dissemelhança).

IV. Exemplo 4

No exemplo em estudo, o conjunto é constituído por 3 classes obtidas segundo os seguintes parâmetros:

Classe A : $\mu_1 = 0$; $\mu_2 = 0$; $\sigma_1^2 = 1.2$; $\sigma_{12} = 1.5$; $\sigma_2^2 = 2$;

Classe B : $\mu_1 = 2$; $\mu_2 = 5$; $\sigma_1^2 = 0.9$; $\sigma_{12} = 0$; $\sigma_2^2 = 0.9$;

Classe C : $\mu_1 = -3$; $\mu_2 = 2$; $\sigma_1^2 = 1$; $\sigma_{12} = -1$; $\sigma_2^2 = 1.2$;

e cuja representação gráfica se encontra na figura 5.4. Comparativamente ao conjunto analisado no Exemplo 2, aproximam-se as classes e alteram-se os valores da matriz de covariâncias por forma a diminuir a compacidade das mesmas.

A tabela 5.9 mostra o valor da média e desvio-padrão do IRM para as 100 réplicas do exemplo em análise. O método NIIVC foi o que produziu melhores resultados, ainda que mais uma vez apresente um elevado desvio-padrão. Logo de seguida, surge o método NIIMEAN e o método das k -médias.

Na amostra em estudo, a tabela 5.10 mostra a vantagem do uso dos valores da função de distribuição da estatística face à aplicação directa da estatística, como

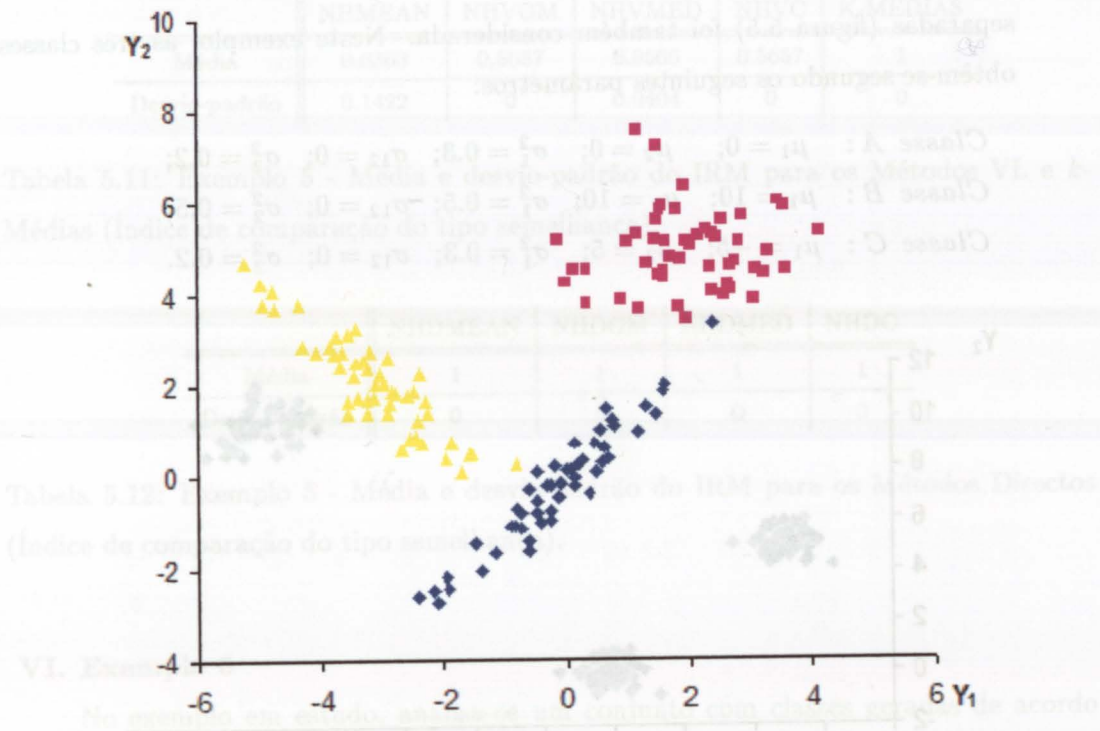


Figura 5.4: Conjunto de dados do Exemplo 4.

	NHMEAN	NHVGCM	NHVMED	NHVC	K-MÉDIAS
Média	0.7930	0.7658	0.7256	0.8520	0.7834
Desvio-padrão	0.0989	0.1287	0.1050	0.1503	0.0854

Tabela 5.9: Exemplo 4 - Média e desvio-padrão do IRM para os Métodos VL e *k*-Médias (Índice de comparação do tipo semelhança).

critério de afectação.

	NHDMEAN	NHDGCM	NHDMED	NHDC
Média	0.6779	0.6774	0.7235	0.6454
Desvio-padrão	0.1195	0.1203	0.1051	0.1030

Tabela 5.10: Exemplo 4 - Média e desvio-padrão do IRM para os Métodos Directos (Índice de comparação do tipo semelhança).

V. Exemplo 5

A análise de conjuntos formados apenas por classes esféricas, compactas e bem

separadas (figura 5.5) foi também considerada. Neste exemplo, as três classes obtêm-se segundo os seguintes parâmetros:

Classe A: $\mu_1 = 0$; $\mu_2 = 0$; $\sigma_1^2 = 0.3$; $\sigma_{12} = 0$; $\sigma_2^2 = 0.2$;

Classe B: $\mu_1 = 10$; $\mu_2 = 10$; $\sigma_1^2 = 0.5$; $\sigma_{12} = 0$; $\sigma_2^2 = 0.5$;

Classe C: $\mu_1 = -5$; $\mu_2 = 5$; $\sigma_1^2 = 0.3$; $\sigma_{12} = 0$; $\sigma_2^2 = 0.2$.

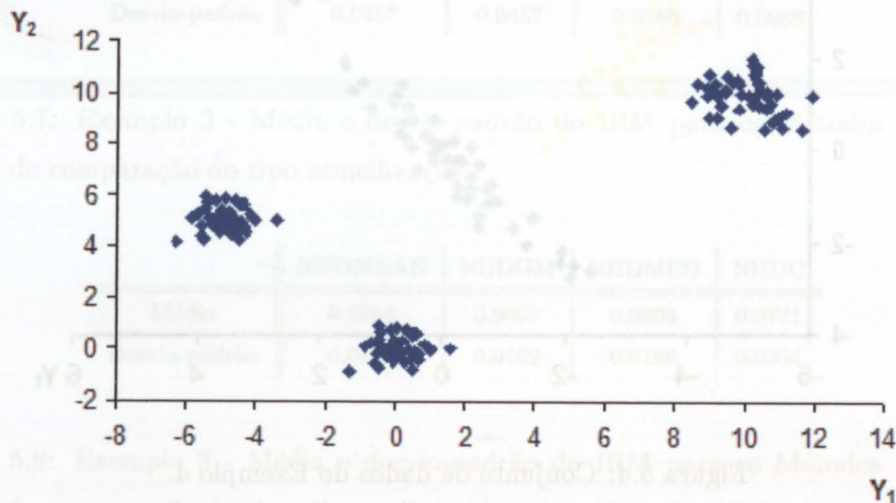


Figura 5.5: Conjunto de dados do Exemplo 5.

Analisando os valores do IRM da tabela 5.11, verifica-se que o método das k -médias e o método NHVMED fornecem bons resultados. Nos métodos NHVGM, NHVC e NHMEAN verificou-se o "efeito de massificação" já observado no método AVD (Agregação pela Validade do Diâmetro) de Classificação Hierárquica (Nicolau [29]). Este efeito baseia-se na formação de classes numerosas: considerando a real estrutura dos dados, sendo I_i um elemento da classe B, a capacidade de atracção da classe A, à qual é atribuído I_i deve-se à cardinalidade da classe A, e é acentuada pelo facto do valor $u(\{I_i\}, A)$, concretização da estatística U subjacente ao método, ser superior a 0.5, tornando o valor da função de distribuição próximo de 1.

Com a aplicação dos métodos directos obtém-se uma partição que separa claramente as três classes que constituem o conjunto (tabela 5.12). De facto, a aplicação directa das estatísticas sobre o conjunto de valores que medem a proximidade entre elementos mostra que o efeito de "massificação" se sobrepõe à real estrutura dos dados.

	NHMEAN	NHVGM	NHVMED	NHVC	K-MÉDIAS
Média	0.6363	0.5657	0.9566	0.5657	1
Desvio-padrão	0.1422	0	0.0404	0	0

Tabela 5.11: Exemplo 5 - Média e desvio-padrão do IRM para os Métodos VL e k -Médias (Índice de comparação do tipo semelhança).

	NHDMEAN	NHDGM	NHDMED	NHDC
Média	1	1	1	1
Desvio-padrão	0	0	0	0

Tabela 5.12: Exemplo 5 - Média e desvio-padrão do IRM para os Métodos Directos (Índice de comparação do tipo semelhança).

VI. Exemplo 6

No exemplo em estudo, analisa-se um conjunto com classes geradas de acordo com os seguintes parâmetros:

Classe A : $\mu_1 = 0$; $\mu_2 = 0$; $\sigma_1^2 = 0.3$; $\sigma_{12} = 0$; $\sigma_2^2 = 0.2$;
Classe B : $\mu_1 = 10$; $\mu_2 = 10$; $\sigma_1^2 = 1$; $\sigma_{12} = 0.7$; $\sigma_2^2 = 0.5$;
Classe C : $\mu_1 = -5$; $\mu_2 = 5$; $\sigma_1^2 = 0.5$; $\sigma_{12} = -0.8$; $\sigma_2^2 = 1$;

e cuja representação gráfica é apresentada na figura 5.6. Relativamente ao Exemplo 5 mantém-se o grau de separação/isolamento das classes, e altera-se a forma esférica das classes B e C para classes alongadas.

	NHMEAN	NHVGM	NHVMED	NHVC	K-MÉDIAS
Média	0.6873	0.5651	0.8576	0.5654	0.9960
Desvio-padrão	0.0996	0.0011	0.0769	0.0008	0.0102

Tabela 5.13: Exemplo 6 - Média e desvio-padrão do IRM para os Métodos VL (Índice de comparação do tipo semelhança).

A tabela 5.13 mostra que, analogamente ao que ocorreu com o conjunto estudado no Exemplo 5, nos métodos NHMEAN, NHVGM e NHVC, verifica-se o efeito de "massificação". De facto, nestes três métodos, o poder de atracção da classe que tem maior cardinalidade é reforçado, ainda que um elemento esteja

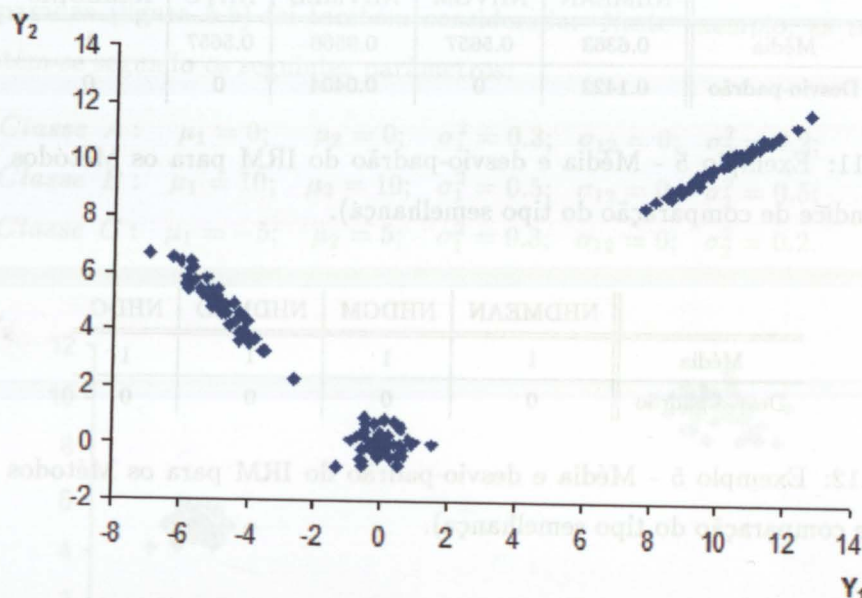


Figura 5.6: Conjunto de dados do Exemplo 6.

mais próximo de outra classe, no sentido da estatística adoptada. Como já foi referido, o elemento será afectado à classe de maior dimensão, o que se acentua quando o valor da estatística sobre o conjunto dos valores de comparação entre o elemento a afectar e os elementos da classe é superior a 0.5. O método NHVMED não sendo afectado por este aspecto, fornece em média melhores resultados.

	NHDMEAN	NHDGM	NHDMED	NHDC
Média	0.9948	0.9835	0.9916	0.9928
Desvio-padrão	0.0098	0.0104	0.0115	0.0132

Tabela 5.14: Exemplo 6 - Média e desvio-padrão do IRM para os Métodos Directos (Índice de comparação do tipo semelhança).

O comportamento dos métodos directos no conjunto de dados em análise é bastante satisfatório (tabela 5.14), mostrando mais uma vez que o efeito de massificação se deve à forte influência da cardinalidade da classe na função de distribuição.

VII. Exemplo 7

Por último, procedeu-se à classificação de um conjunto cujas classes apresentam um elevado grau de separação, mantendo-se os parâmetros de correlação definidos no Exemplo 5:

Classe A : $\mu_1 = 10; \mu_2 = -20; \sigma_1^2 = 0.3; \sigma_{12} = 0; \sigma_2^2 = 0.2;$

Classe B : $\mu_1 = 10; \mu_2 = 10; \sigma_1^2 = 0.5; \sigma_{12} = 0; \sigma_2^2 = 0.5;$

Classe C : $\mu_1 = -24; \mu_2 = -24; \sigma_1^2 = 0.3; \sigma_{12} = 0; \sigma_2^2 = 0.2.$

A figura 5.7 mostra a representação gráfica do conjunto em análise.

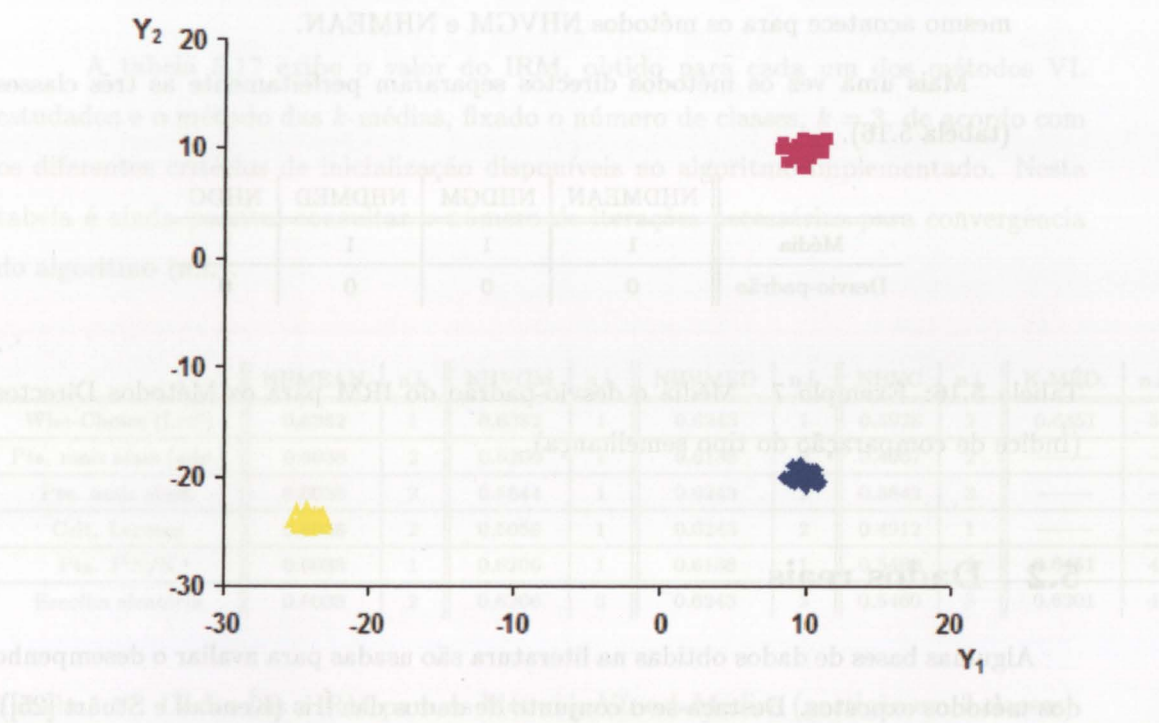


Figura 5.7: Conjunto de dados do Exemplo 7.

	NHMEAN	NHVGCM	NHVMED	NHVC	K-MÉDIAS
Média	1	1	1	1	1
Desvio-padrão	0	0	0	0	0

Tabela 5.15: Exemplo 7 - Média e desvio-padrão do IRM para os Métodos VL e *k*-Médias (Índice de comparação do tipo semelhança).

Contrariamente ao que se verificou nos dois exemplos anteriores, neste con-

junto qualquer um dos métodos separa perfeitamente as três classes. Uma análise mais cuidada sobre os resultados obtidos permite verificar que na classificação efectuada pelo método NHVC obtida uma das classes, por exemplo a classe A, constituída pelos elementos I_1 a I_{50} , no momento da classificação do elemento I_ℓ , $50 < \ell < 101$, elemento pertencente à classe B, o valor da estatística centro das semelhanças VL entre o indivíduo I_ℓ e os indivíduos pertencentes à classe B, $u_C(\{I_\ell\}, B)$, é suficientemente grande, não permitindo que a elevada cardinalidade da classe A aumente o valor de $\Gamma_{NHVC}(u_C(\{I_\ell\}, A))$ e este seja superior a $\Gamma_{NHVC}(u_C(\{I_\ell\}, B))$, atribuindo-se correctamente o elemento I_ℓ à classe B. O mesmo acontece para os métodos NHVGM e NHMEAN.

Mais uma vez os métodos directos separaram perfeitamente as três classes (tabela 5.16).

	NHDMEAN	NHDGM	NHDMED	NHDC
Média	1	1	1	1
Desvio-padrão	0	0	0	0

Tabela 5.16: Exemplo 7 - Média e desvio-padrão do IRM para os Métodos Directos (Índice de comparação do tipo semelhança).

5.2 Dados reais

Algumas bases de dados obtidas na literatura são usadas para avaliar o desempenho dos métodos expostos. Destaca-se o conjunto de dados das Íris (Kendall e Stuart [25]), no qual o estudo é mais exaustivo.

5.2.1 Dados das Íris

Os dados aqui tratados consistem num conjunto de 150 flores de tipo Íris, 50 da variedade Setosa, 50 da variedade Versicolor e 50 do tipo Virgínica, descritas por quatro variáveis: comprimento da sépala, largura da sépala, comprimento da pétala e largura da pétala. Numerados os elementos de 1 a 150, os primeiros 50 indivíduos, I_1 a I_{50} , correspondem à variedade Íris Setosa, de I_{51} a I_{100} tem-se a variedade Íris Versicolor e os restantes correspondem à variedade Virgínica, formando assim a partição referência:

- Classe A: $I_1 - I_{50}$; Classe B: $I_{51} - I_{100}$; Classe C: $I_{101} - I_{150}$.

Relativamente à escolha do índice de comparação entre elementos foram testados dois índices do tipo semelhança:

$$s_1(I_i, I_\ell) = \frac{1 + \cos \alpha}{2},$$

(5.2.1)

sendo α o ângulo formado pelos indivíduos I_i e I_ℓ , e

$$s_2(I_i, I_\ell) = \frac{1}{1 + d(I_i, I_\ell)},$$

sendo $d(I_i, I_\ell)$ a distância euclideana entre I_i e I_ℓ . Nas experiências realizadas, o índice s_1 forneceu melhores resultados para o IRM pelo que se optou por usar a semelhança s_1 .

A tabela 5.17 exhibe o valor do IRM, obtido para cada um dos métodos VL estudados e o método das k -médias, fixado o número de classes, $k = 3$, de acordo com os diferentes critérios de inicialização disponíveis no algoritmo implementado. Nesta tabela é ainda possível consultar o número de iterações necessárias para convergência do algoritmo (n.i.).

	NHMEAN	n.i.	NHVGM	n.i.	NHVMED	n.i.	NHVC	n.i.	K-MÉD.	n.i.
Wise-Choice (L=5)	0.6352	1	0.6352	1	0.6243	1	0.5928	2	0.6451	5
Pts. mais afast.(adp.)	0.6038	2	0.6206	1	0.6138	2	0.4967	2	—	—
Pts. mais afast.	0.6038	2	0.5844	1	0.6243	2	0.5842	3	—	—
Crit. Lerman	0.6038	2	0.5058	1	0.6243	2	0.4912	1	—	—
Pts. J*N/K	0.6038	1	0.6206	1	0.6138	1	0.5460	2	0.6451	4
Escolha aleatória	0.6038	2	0.6206	3	0.6243	2	0.5460	3	0.6201	4

Tabela 5.17: Dados Íris - IRM para os Métodos VL e k -Médias (partição em 3 classes).

Analisando os resultados expressos na tabela 5.17 verifica-se que o método *wise-choice*, com as adaptações referidas na Secção 3.3, é em geral um bom critério de escolha dos centros iniciais, salientando-se o número reduzido de iterações necessário para atingir a convergência. Tomando o critério de inicialização "pontos mais afastados", com as adaptações sugeridas em 3.3, neste conjunto o método NHVC não se revelou eficaz, dado que nos centros iniciais fornecidos constam dois pontos que pertencem à mesma classe (I_{34} e I_{42}). O baixo valor do IRM deve-se ao facto de uma das classes da partição final ser constituída por apenas dois elementos. Pelo contrário, este tipo de inicialização nos métodos NHVGM e NHVMED teve melhor desempenho.

Considerando o mesmo critério "pontos mais afastados", sem as adaptações sugeridas, é o método NHVMED que fornece melhores resultados.

A tabela 5.18 mostra a partição solução fornecida por cada um dos métodos VL e o método das k -médias de MacQueen, tendo por base o critério de inicialização *wise-choice*. Os métodos NHVMED, NHVGM e NIIMEAN, agrupam numa única classe os primeiros 50 elementos (espécie Íris Setosa), verificando-se uma mistura de elementos das outras duas espécies nas restantes classes. A percentagem de elementos mal classificados nos métodos NHVGM e NIIMEAN é de 15.3%, no método NHVMED é de 16.7%, no método das k -médias é de 14.7% e no método NHVC é de 17.3%.

As tabelas 5.19, 5.20, 5.21 e 5.22 avaliam a homogeneidade e isolamento das

	Classe A	Classe B	Classe C
NHVMED	$I_1 - I_{50}$	$I_{51} - I_{53}; I_{55}; I_{57}; I_{59}; I_{62}; I_{64}; I_{66}$ $I_{71}; I_{75} - I_{79}; I_{86}; I_{87}; I_{92}$ $I_{98}; I_{101}; I_{103} - I_{106}; I_{108} - I_{113}$ $I_{115} - I_{119}; I_{121}; I_{123} - I_{146}; I_{148} - I_{150}$	$I_{54}; I_{56}; I_{58}; I_{60}; I_{61}; I_{63}; I_{65}$ $I_{67} - I_{70}; I_{72} - I_{74}; I_{80} - I_{85}$ $I_{88} - I_{91}; I_{93} - I_{97}; I_{99}; I_{100}$ $I_{102}; I_{107}; I_{114}; I_{120}; I_{122}; I_{147}$
NHVGM	$I_1 - I_{50}$	$I_{51} - I_{53}; I_{55}; I_{57}; I_{59}; I_{62}; I_{64}$ $I_{66}; I_{71}; I_{75} - I_{79}; I_{86}; I_{87}$ $I_{92}; I_{98}; I_{101}; I_{103} - I_{106}$ $I_{108} - I_{113}; I_{115} - I_{119}; I_{121}; I_{123} - I_{150}$	$I_{54}; I_{56}; I_{58}; I_{60}; I_{61}; I_{63}; I_{65}$ $I_{67} - I_{70}; I_{72} - I_{74}; I_{80} - I_{85}$ $I_{88} - I_{91}; I_{93} - I_{97}; I_{99}; I_{100}$ $I_{102}; I_{107}; I_{114}; I_{120}; I_{122}$
NHVC	$I_1 - I_{50}; I_{58}$ $I_{60}; I_{65}$	$I_{51} - I_{53}; I_{55}; I_{57}; I_{59}; I_{62}$ $I_{64}; I_{66}; I_{67}; I_{71}; I_{75} - I_{79}$ $I_{86}; I_{87}; I_{92}; I_{101}; I_{103} - I_{106}; I_{108}$ $I_{110} - I_{113}; I_{115} - I_{119}; I_{121} - I_{150}$	$I_{54}; I_{56}; I_{61}; I_{63}$ $I_{68} - I_{70}; I_{72} - I_{74}; I_{80} - I_{85}$ $I_{88} - I_{91}; I_{93} - I_{100}; I_{102}$ $I_{107}; I_{109}; I_{114}; I_{120}$
NIIMEAN	$I_1 - I_{50}$	$I_{51} - I_{53}; I_{55}; I_{57}; I_{59}; I_{62}; I_{64}$ $I_{66}; I_{71}; I_{75} - I_{79}; I_{86}; I_{87}$ $I_{92}; I_{98}; I_{101}; I_{103} - I_{106}$ $I_{108} - I_{113}; I_{115} - I_{119}; I_{121}; I_{123} - I_{150}$	$I_{54}; I_{56}; I_{58}; I_{60}; I_{61}; I_{63}; I_{65}$ $I_{67} - I_{70}; I_{72} - I_{74}; I_{80} - I_{85}$ $I_{88} - I_{91}; I_{93} - I_{97}; I_{99}; I_{100}$ $I_{102}; I_{107}; I_{114}; I_{120}; I_{122}$
K-MÉDIAS	$I_1 - I_{41}$ $I_{43} - I_{50}$	$I_{42}; I_{54}; I_{56}; I_{58}; I_{60} - I_{65}$ $I_{67} - I_{70}; I_{72} - I_{75}; I_{79} - I_{85}$ $I_{88} - I_{100}; I_{102}; I_{107}; I_{114}$ $I_{120}; I_{122}; I_{135}; I_{143}; I_{147}$	$I_{51} - I_{53}; I_{55}; I_{57}; I_{59}; I_{66}; I_{71}$ $I_{76} - I_{78}; I_{86}; I_{87}; I_{101}; I_{103} - I_{106}$ $I_{108} - I_{113}; I_{115} - I_{119}; I_{121}; I_{123} - I_{134}$ $I_{136} - I_{142}; I_{144} - I_{146}; I_{148} - I_{150}$

Tabela 5.18: Dados Íris - Partição em 3 classes obtidas pelos Métodos VL e k -Médias (com inicialização *wise-choice*).

classes obtidas com os métodos VL e que figuram na tabela 5.18. A notação usada para cada uma das medidas é a mesma do Capítulo 4.

NHVC	H_2	H_3	H_5	H_4	H_6	I_{so1}	I_{so2}
Classe A	0.1629	0.8589	0.7928	0.8022	0.5392	0.9141	0.2091
Classe B	0.4203	0.8681	0.8537	0.8448	0.6678	0.9155	0.3179
Classe C	0.4499	0.8451	0.8474	0.8244	0.6827	0.9155	0.4641

Tabela 5.19: Dados Íris - Medidas de homogeneidade e isolamento das classes obtidas pelo método NHVC.

NHVGM	H_2	H_3	H_9	H_4	H_{10}	I_{so1}	I_{so2}
Classe A	0.3820	0.8799	0.1082	0.8412	0.1338	0.8940	0.1992
Classe B	0.4233	0.8713	0.1225	0.8512	0.1646	0.9155	0.3139
Classe C	0.5068	0.8493	0.1359	0.8230	0.2010	0.9155	0.4512

Tabela 5.20: Dados Íris - Medidas de homogeneidade e isolamento das classes obtidas pelo método NHVGM.

NHVMED	H_2	H_3	H_7	H_4	H_8	I_{so1}	I_{so2}
Classe A	0.3820	0.8799	0.9072	0.8412	0.8969	0.8940	0.1992
Classe B	0.4233	0.8717	0.8928	0.8526	0.8748	0.9155	0.3182
Classe C	0.5068	0.8494	0.8866	0.8218	0.8498	0.9155	0.4496

Tabela 5.21: Dados Íris - Medidas de homogeneidade e isolamento das classes obtidas pelo método NHVMED.

NHMEAN	H_2	H_3	H_4	I_{so1}	I_{so2}
Classe A	0.3820	0.8799	0.8410	0.8940	0.1992
Classe B	0.4233	0.8713	0.8512	0.9155	0.3139
Classe C	0.5068	0.8493	0.8230	0.9155	0.4512

Tabela 5.22: Dados Íris - Medidas de homogeneidade e isolamento das classes obtidas pelo método NHMEAN.

A medida de homogeneidade "Diâmetro", H_2 , dada pelo valor da semelhança entre os dois elementos mais afastados da classe, sugere que a classe mais heterogénea é a classe A, sendo a classe C a classe mais homogénea. Baseada no máximo da média das semelhanças entre um elemento da classe e os restantes elementos da mesma classe, a medida "Estrela", H_3 , identifica a classe A, para os métodos NHVMED, NHVGM e

NHMEAN, como a classe mais homogénea. Já a medida de homogeneidade H_4 , que se obtém somando os valores das semelhanças da matriz triangular inferior, sugere que a classe B é a mais homogénea, para qualquer um dos métodos VL. As medidas de isolamento I_{so1} e I_{so2} , definidas em (4.2.1) e (4.2.2), respectivamente, mostram que a classe mais isolada da partição é a classe A.

A avaliação global das partições de acordo com as medidas G_3 , G_5 e G_6 , que usam a função de distribuição da estatística média aritmética, mediana e média geométrica, respectivamente, sobre as semelhanças VL entre todos os pares de elementos pertencentes à mesma classe (Secção 4.3) atribui o valor 1 às partições obtidas, à excepção da medida G_4 , dada pela expressão em (4.3.2), valor da função de distribuição do centro das semelhanças VL entre pares de elementos da mesma classe, que atribui o valor 0 à partição obtida pelo método NHVC. O valor 1 atribuído à partição dada pelos métodos NHMEAN, NHVMED e NHVGM, deve-se ao elevado número de pares de elementos pertencentes à mesma classe, aliado ao facto de o valor da estatística das semelhanças VL entre todos os pares de elementos pertencentes à mesma classe ser superior a 0.5. Por sua vez, o valor atribuído à partição sugerida pelo método NHVC deve-se igualmente ao elevado número de pares de elementos da mesma classe e ao facto do centro das semelhanças entre todos os pares de elementos pertencentes à mesma classe ser inferior a 0.5.

Com a aplicação das várias medidas de homogeneidade e de isolamento de uma classe expostas na Secção 4.1 pretendia-se identificar a classe mais homogénea e isolada das partições obtidas pelos diferentes métodos, não sendo o estudo realizado conclusivo. Note-se ainda que nenhum estudo foi feito no sentido de analisar se os métodos de classificação propostos optimizam os critérios de homogeneidade e isolamento.

Dado o número máximo de três classes e permitindo a fusão de classes, todos os métodos VL sugerem uma partição solução em duas classes (tabela 5.23). Em todos os métodos, a partição solução agrupa na mesma classe os primeiros 50 elementos da variedade Íris Setosa e um reduzido número de elementos da variedade Versicolor; no caso dos métodos NHVMED e NHMEAN apenas o elemento I_{58} pertence a esta mesma classe.

	Classe A	Classe B
NHVC	$I_1 - I_{50}$ $I_{58}; I_{60}; I_{61}; I_{65}; I_{80}; I_{82}; I_{94}; I_{99}$	$I_{51} - I_{57}; I_{59}; I_{62} - I_{64}; I_{66} - I_{79}$ $I_{81}; I_{83} - I_{93}; I_{95} - I_{98}; I_{100} - I_{150}$
NHVGGM	$I_1 - I_{50}$ $I_{58}; I_{60}; I_{61}; I_{65}; I_{94}; I_{99}$	$I_{51} - I_{57}; I_{59}; I_{62} - I_{64}; I_{66} - I_{93}$ $I_{95} - I_{98}; I_{100} - I_{150}$
NHVMED	$I_1 - I_{50}; I_{58}$	$I_{51} - I_{57}; I_{59} - I_{150}$
NHMEAN	$I_1 - I_{50}; I_{58}$	$I_{51} - I_{57}; I_{59} - I_{150}$

Tabela 5.23: Dados Íris - Partição em 2 classes obtida pelos Métodos VL.

	NHDMEAN		NHDGM		NHDMED		NHDC	
Wise-Choice ($L=5$)	0.5961	2	0.5794	1	0.5875	1	0.5841	1
Pts. mais afastados	0.3868	1	0.5794	1	0.3976	2	0.4967	2
Crit. Lerman	0.3868	2	0.3729	1	0.3976	1	0.6294	4
Pts. J*N/K	0.5875	1	0.5875	1	0.5875	2	0.6044	1
Escolha aleatória	0.4005	3	0.5875	2	0.5875	1	0.5694	4

Tabela 5.24: Dados Íris - IRM para os Métodos Directos (partição em 3 classes).

O conjunto em análise foi classificado de acordo com os Métodos Directos abordados na Secção 3.2, considerando uma partição em três classes. A tabela 5.24 exhibe os valores do IRM para os diferentes critérios de inicialização. A comparação destes valores com os obtidos aplicando a metodologia VL (tabela 5.17), permite concluir que o uso da função de distribuição da estatística em alternativa à aplicação directa da estatística subjacente ao método fornece melhores resultados.

Na aplicação dos métodos directos, com o critério de inicialização *wise-choice*, para $L = 5$, verificou-se que a média aritmética e a mediana são as estatísticas de tendência central que têm melhor desempenho. Nesta classificação é ainda possível verificar que embora o elemento I_{42} seja um elemento que esteja incluído na classe A da partição referência, a sua L -vizinhança indica que este ponto se encontra mais próximo de elementos da classe B, facto ao qual as estatísticas média aritmética, média geométrica e mediana são mais sensíveis. Por sua vez, o método directo que tem por base o centro amostral, NHDC, agrupa este elemento na primeira classe.

5.2.2 Dados PHYTOS

A informação disponível neste conjunto de dados binários consta de 16 parcelas de terreno caracterizadas pela presença ou ausência de 66 espécies vegetais (Roux [36]). Assim, analisa-se um conjunto de 16 indivíduos que são descritos por 66 variáveis. Toma-se como partição referência uma partição sugerida por Roux [36].

Usa-se o índice de semelhança VL para dados binários que tem por base a escolha do índice expresso em (1.3.1). Na tabela 5.25, regista-se o valor do IRM de acordo com diferentes critérios de inicialização para uma partição em quatro classes e na tabela 5.26 registam-se os resultados da classificação, usando o critério de inicialização *wise-choice*, com $L = 5$.

	NHMEAN		NHVGM		NHVMED		NHVC	
Wise-Choice (L=5)	0.5180	2	0.5418	1	0.5180	2	0.6644	1
Pts. mais afastados	0.5180	1	0.5180	1	0.5180	1	0.4262	1
Crit. Lerman	0.5180	1	0.5180	3	0.5180	1	0.5180	1
Pts. J*N/K	0.5522	1	0.3534	1	0.3344	1	0.5522	1
Escolha aleatória	0.4251	1	0.5180	1	0.5302	2	0.5884	1

Tabela 5.25: Dados Phytos - IRM obtidos pelos Métodos VL (partição em 4 classes).

	Classe A	Classe B	Classe C	Classe D
Part. de Roux	$I_4; I_6; I_8$	$I_1; I_2; I_5; I_7; I_9$	$I_3; I_{15}; I_{16}$	$I_{10} - I_{14}$
NHMEAN	$I_1 - I_3; I_5; I_7; I_9; I_{16}$	$I_{11}; I_{12}; I_{14}$	$I_4; I_6; I_8; I_{15}$	$I_{10}; I_{13}$
NHVGM	$I_1; I_2; I_4 - I_9$	$I_{11}; I_{12}; I_{14}$	$I_3; I_{15}; I_{16}$	$I_{10}; I_{13}$
NHVMED	$I_1 - I_3; I_5; I_7; I_9; I_{16}$	$I_{11}; I_{12}; I_{14}$	$I_4; I_6; I_8; I_{15}$	$I_{10}; I_{13}$
NHVC	$I_1 - I_3; I_5 - I_7; I_9$	$I_{10} - I_{14}$	$I_4; I_8$	$I_{15}; I_{16}$

Tabela 5.26: Dados Phytos - Partição em 4 classes obtida pelos Métodos VL (com critério de inicialização *wise-choice*).

Analisando os valores que constam na tabela 5.25, verifica-se que em geral o critério *wise-choice* continua a ser uma boa opção de inicialização dos centros. Além disso, salienta-se a rápida convergência para a partição solução em qualquer um dos métodos.

Uma análise sobre a tabela 5.26, mostra a igualdade entre a partição solução obtida

com os métodos NHMEAN e NHVMED.

A aplicação dos métodos directos, fixado o número de classes, $k = 4$, conduziu aos resultados expressos na tabela 5.27. Considerando o critério de inicialização *wise-choice* verifica-se que é mais vantajoso o uso dos métodos NHVGM e NHVC do que os respectivos métodos directos. De salientar os valores do IRM elevados que foram encontrados com os métodos NHDMEAN e NHDGM quando combinados com o critério de inicialização "escolha aleatória".

	NHDMEAN		NHDGM		NHDMED		NHDC	
Wise-Choice (L=5)	0.5180	1	0.5400	1	0.5400	1	0.3344	1
Pts. J*N/K	0.5522	1	0.5522	1	0.3344	1	0.5522	1
Escolha aleatória	0.8257	2	0.7261	1	0.4745	2	0.4440	2

Tabela 5.27: Dados Phytos - IRM obtidos pelos Métodos Directos (partição em 4 classes).

5.2.3 Dados Ruspini

A figura 5.8 ilustra os dados de Ruspini (cf. Kaufman e Rousseeuw [24]). Este conjunto é formado por 75 elementos descritos por 2 variáveis quantitativas. A tabela 5.28 indica as coordenadas dos pontos em estudo, numerados de 1 a 75.

Os pontos são escolhidos tendo em conta a formação de 4 classes distintas. A partição de referência usada é:

- Classe A: $I_1 - I_{20}$; Classe B: $I_{21} - I_{43}$; Classe C: $I_{44} - I_{60}$; Classe D: $I_{61} - I_{75}$.

O índice de comparação entre elementos é novamente o índice expresso em (5.2.1).

A tabela 5.29 apresenta os resultados obtidos, para um número fixo de 4 classes. No método das k -médias, os centros iniciais são os pontos de ordem $\frac{j \times n}{k}$, $j = 1, \dots, k$, e para os métodos VL a inicialização recai sobre o critério *wise-choice*, com $L = 5$.

Mais uma vez, salienta-se o baixo número de iterações necessário para atingir a convergência em cada um dos métodos, uma vez que a partição solução é obtida logo na primeira iteração, com excepção do método NHVC que exige mais uma iteração. Ainda relativamente a este método, acrescenta-se que a partição obtida na primeira iteração apresenta um valor do IRM superior ao valor apresentado pela partição fornecida no momento de convergência do algoritmo, o que mostra que neste caso, o critério usado

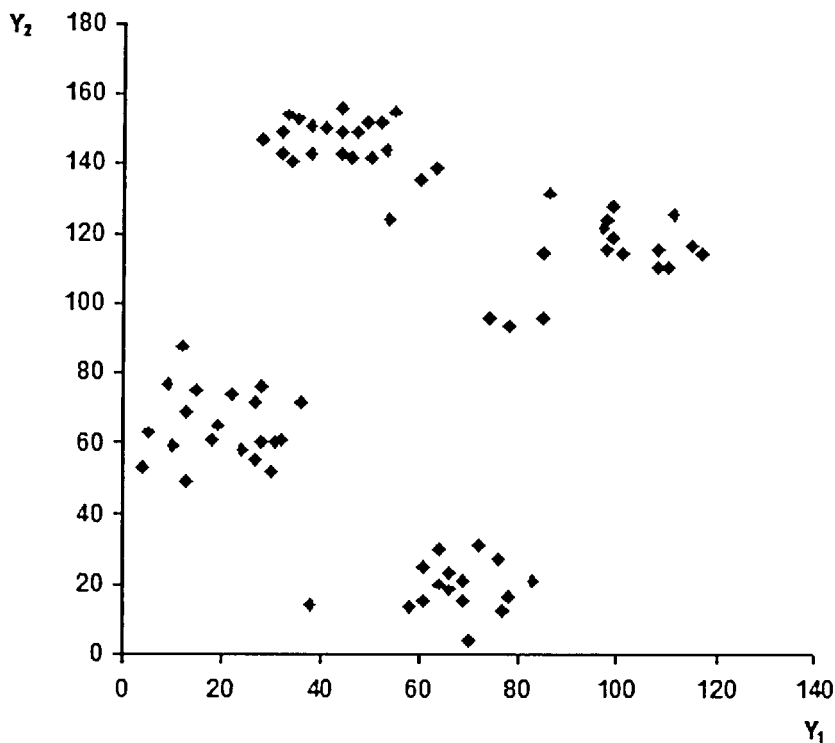


Figura 5.8: Dados Ruspini

para a definição do centro de cada classe não é o mais adequado. O método NHVMED é o método VL que melhor se comporta neste conjunto de dados. Note-se que o baixo valor do IRM nos métodos NHVGM e NIHC deve-se ao efeito de massificação já verificado noutros exemplos.

A homogeneidade e isolamento das classes da partição solução obtidas pelos novos métodos VL propostos, foram avaliadas tendo em conta algumas das medidas referidas no Capítulo 4 (tabelas 5.30, 5.31 e 5.32).

Analisando a tabela 5.30 verifica-se que as medidas de homogeneidade "Diâmetro", H_2 , e "Estrela", H_3 , avaliam a classe D como a menos homogênea e a classe A como a mais homogênea. A medida H_4 , "Soma das Distâncias", avalia a classe B como a mais homogênea. Analogamente às medidas H_2 e H_3 , H_4 sugere a classe D como a classe menos homogênea. A medida H_9 sugere a classe A como a classe mais homogênea. Analisando o isolamento, a medida Iso_1 avalia a classe A como a mais isolada, e a medida Iso_2 considera a classe D a mais isolada.

	Y ₁	Y ₂		Y ₁	Y ₂		Y ₁	Y ₂		Y ₁	Y ₂		Y ₁	Y ₂
<i>I</i> ₁	4	53		<i>I</i> ₁₆	28 60		<i>I</i> ₃₁	44 156		<i>I</i> ₄₆	85 96		<i>I</i> ₆₁	70 4
<i>I</i> ₂	5	63		<i>I</i> ₁₇	30 52		<i>I</i> ₃₂	44 149		<i>I</i> ₄₇	78 94		<i>I</i> ₆₂	77 12
<i>I</i> ₃	10	59		<i>I</i> ₁₈	31 60		<i>I</i> ₃₃	44 143		<i>I</i> ₄₈	74 96		<i>I</i> ₆₃	83 21
<i>I</i> ₄	9	77		<i>I</i> ₁₉	32 61		<i>I</i> ₃₄	46 142		<i>I</i> ₄₉	97 122		<i>I</i> ₆₄	61 15
<i>I</i> ₅	13	49		<i>I</i> ₂₀	36 72		<i>I</i> ₃₅	47 149		<i>I</i> ₅₀	98 116		<i>I</i> ₆₅	69 15
<i>I</i> ₆	13	69		<i>I</i> ₂₁	28 147		<i>I</i> ₃₆	49 152		<i>I</i> ₅₁	98 124		<i>I</i> ₆₆	78 16
<i>I</i> ₇	12	88		<i>I</i> ₂₂	32 149		<i>I</i> ₃₇	50 142		<i>I</i> ₅₂	99 119		<i>I</i> ₆₇	66 18
<i>I</i> ₈	15	75		<i>I</i> ₂₃	35 153		<i>I</i> ₃₈	53 144		<i>I</i> ₅₃	99 128		<i>I</i> ₆₈	58 13
<i>I</i> ₉	18	61		<i>I</i> ₂₄	33 154		<i>I</i> ₃₉	52 152		<i>I</i> ₅₄	101 115		<i>I</i> ₆₉	64 20
<i>I</i> ₁₀	19	65		<i>I</i> ₂₅	38 151		<i>I</i> ₄₀	55 155		<i>I</i> ₅₅	108 111		<i>I</i> ₇₀	69 21
<i>I</i> ₁₁	22	74		<i>I</i> ₂₆	41 150		<i>I</i> ₄₁	54 124		<i>I</i> ₅₆	110 111		<i>I</i> ₇₁	66 23
<i>I</i> ₁₂	27	72		<i>I</i> ₂₇	38 145		<i>I</i> ₄₂	60 136		<i>I</i> ₅₇	108 116		<i>I</i> ₇₂	61 25
<i>I</i> ₁₃	28	76		<i>I</i> ₂₈	38 143		<i>I</i> ₄₃	63 139		<i>I</i> ₅₈	111 126		<i>I</i> ₇₃	76 27
<i>I</i> ₁₄	24	58		<i>I</i> ₂₉	32 143		<i>I</i> ₄₄	86 132		<i>I</i> ₅₉	115 117		<i>I</i> ₇₄	72 31
<i>I</i> ₁₅	27	55		<i>I</i> ₃₀	34 141		<i>I</i> ₄₅	85 115		<i>I</i> ₆₀	117 115		<i>I</i> ₇₅	64 30

Tabela 5.28: Dados Ruspini

	NHVGM	NHVC	NHVMED	NHMEAN	K-MÉDIAS
Classe A	<i>I</i> ₆₁ – <i>I</i> ₇₅	<i>I</i> ₁ – <i>I</i> ₂₀ ; <i>I</i> ₆₁ – <i>I</i> ₇₅	<i>I</i> ₁ – <i>I</i> ₂₀	<i>I</i> ₆₁ – <i>I</i> ₇₅	<i>I</i> ₁ – <i>I</i> ₂₀
Classe B	<i>I</i> ₃₁ ; <i>I</i> ₄₃ – <i>I</i> ₅₈ ; <i>I</i> ₆₀	<i>I</i> ₄₃	<i>I</i> ₄₅ – <i>I</i> ₆₀	<i>I</i> ₂₂ – <i>I</i> ₄₅	<i>I</i> ₂₁ – <i>I</i> ₄₃
Classe C	<i>I</i> ₅₉	<i>I</i> ₄₇	<i>I</i> ₂₁ – <i>I</i> ₄₄	<i>I</i> ₄₆ – <i>I</i> ₆₀	<i>I</i> ₄₄ – <i>I</i> ₆₀
Classe D	<i>I</i> ₁ – <i>I</i> ₃₀ ; <i>I</i> ₃₂ – <i>I</i> ₄₂	<i>I</i> ₂₁ – <i>I</i> ₄₂ ; <i>I</i> ₄₄ – <i>I</i> ₄₆ <i>I</i> ₄₈ – <i>I</i> ₆₀	<i>I</i> ₆₁ – <i>I</i> ₇₅	<i>I</i> ₁ – <i>I</i> ₂₁	<i>I</i> ₆₁ – <i>I</i> ₇₅
IRM	0.5856	0.4859	0.9623	0.8881	1.000

Tabela 5.29: Dados Ruspini - Partição em 4 classes obtida pelos Métodos VL e *k*-Médias.

NHVGM	<i>H</i> ₂	<i>H</i> ₃	<i>H</i> ₉	<i>H</i> ₄	<i>I</i> _{so1}	<i>I</i> _{so2}
Classe A	0.8955	0.8602	0.0816	0.7966	0.8301	0.3739
Classe B	0.3879	0.8459	0.1142	0.8630	0.9237	0.3752
Classe C	—	—	—	—	0.9237	0.4813
Classe D	0.1233	0.7426	0.2975	0.6646	0.9237	0.3336

Tabela 5.30: Dados Ruspini - Medidas de homogeneidade e isolamento das classes obtidas pelo método NHVGM.

NHVC	H_2	H_3	H_5	H_4	H_6	Iso_1	Iso_2
Classe A	0.2506	0.8132	0.7761	0.7072	0.5871	0.7934	0.2710
Classe B	—	—	—	—	—	0.9230	0.5129
Classe C	—	—	—	—	—	0.9236	0.4740
Classe D	0.2223	0.8073	0.7853	0.7122	0.5730	0.9236	0.2965

Tabela 5.31: Dados Ruspini - Medidas de homogeneidade e isolamento das classes obtidas pelo método NHVC.

NHVMED	H_2	H_3	H_7	H_4	Iso_1	Iso_2
Classe A	0.8575	0.8739	0.9218	0.4119	0.8301	0.3639
Classe B	0.9046	0.8644	0.9229	0.4601	0.9194	0.3795
Classe C	0.5135	0.8707	0.9219	0.4454	0.9194	0.3497
Classe D	0.8955	0.8602	0.9228	0.4597	0.8301	0.3739

Tabela 5.32: Dados Ruspini - Medidas de homogeneidade e isolamento das classes obtidas pelo método NHVMED.

Na avaliação da homogeneidade das classes da partição obtida pelo método NHVC, tem-se que as medidas H_2 , H_3 e H_6 avaliam a classe A como a mais homogênea, e as medidas H_4 e H_5 consideram a classe D a mais homogênea. No que respeita ao isolamento das classes, as duas medidas Iso_1 e Iso_2 consideram a classe A a mais isolada.

Na avaliação das classes obtidas com o método NHVMED, as medidas H_2 , H_7 e H_4 consideram a classe B a mais homogênea, a medida H_3 sugere a classe A como a mais homogênea. Quanto ao isolamento das classes, a medida Iso_1 aponta como classes mais isoladas, as classes A e D, e a medida Iso_2 considera a classe C a mais isolada.

A não existência de valores na avaliação da homogeneidade da classe C, obtida com o método NHVGM, e das classes B e C, obtidas com o método NIIVC, deve-se ao facto destas classes terem cardinal um.

Inicializando o algoritmo com um número máximo de $k = 4$ e permitindo a fusão de classes, a partição solução é constituída por três classes no caso dos métodos NIIVC e NHVMED. Já os métodos NIIMEAN e NHVGM fornecem partições em apenas duas classes (tabela 5.33).

	NHVGGM	NHVC	NHVMED	NHMEAN
Classe A	$I_{31} ; I_{43} - I_{75}$	$I_{61} - I_{75}$	$I_1 - I_{20}; I_{61}; I_{64}; I_{65}; I_{68} - I_{72}; I_{75}$	$I_1 - I_{22}; I_{61} - I_{75}$
Classe B	$I_1 - I_{30} ; I_{32} - I_{42}$	$I_{42} - I_{60}$	$I_{46} - I_{60}; I_{62}; I_{63}; I_{66}; I_{67}; I_{73}; I_{74}$	$I_{23} - I_{60}$
Classe C	—	$I_1 - I_{41}$	$I_{21} - I_{45}$	—

Tabela 5.33: Dados Ruspini - Partição obtida pelos Métodos VL (com fusão de classes).

Este conjunto de dados foi ainda submetido aos Métodos Directos abordados na Secção 3.2. Em qualquer um dos casos a partição solução, obtida logo na primeira iteração, é igual à partição referência considerada.

5.2.4 Dados de Nutrientes

Os dados em estudo foram retirados de "Yearbook of Agriculture, 1959" (Departamento de Agricultura dos Estados Unidos, Washington) (cf. Hartigan [20]). Consiste num conjunto de 27 alimentos (carnes e peixes) caracterizados pela energia, proteína, gordura, cálcio e ferro por cada 3 onças (tabela 5.34). Novamente, a medida de semelhança usada foi a indicada nos exemplos anteriores para dados métricos.

Foram aplicados os métodos VL e k -médias de Classificação Não Hierárquica, e o método de Ward de Classificação Hierárquica. Os resultados obtidos constam nas tabelas 5.35 e 5.36, para um número fixo de 2 e 3 classes, respectivamente. O critério de inicialização é o *wise-choice*, no caso dos métodos VL, e para o método das k -médias os centros iniciais são os pontos da forma $\frac{j \times n}{k}$, $j = 1, \dots, k$.

A partição em duas classes é igual para todos os métodos VL (tabela 5.35), sendo muito semelhante à partição fornecida pelo método das k -médias (o valor do IRM para as duas partições é de 0.7152). A classe A obtida pelos métodos VL agrupa os elementos com elevada energia e gordura. Por outro lado, a classe B contém os alimentos com baixo valor de energia e proteínas, ou alto valor em cálcio.

Para uma partição em três classes (tabela 5.36), os métodos VL continuam a agrupar na mesma classe os alimentos cujo valor de energia e proteínas é elevado, na classe B reúnem-se os elementos que apresentam um alto valor de ferro. Sublinha-se a igualdade da classificação obtida pelos métodos NHVC e NHVMED, que também se encontra muito próxima da classificação obtida pelo método NHVGGM; o valor do IRM de entre

	Descrição dos alimentos	Energia (Kcal)	Proteína (g)	Gordura (g)	Cálcio (mg)	Ferro (mg)
I_1	Carne estufada	340	20	28	9	2.6
I_2	Hamburguer	245	21	17	9	2.7
I_3	Carne/A	420	15	39	7	2
I_4	Bife	375	19	32	9	2.6
I_5	Carne/E	180	22	10	17	3.7
I_6	Galinha/G	115	20	3	8	1.4
I_7	Galinha/E	170	25	7	12	1.5
I_8	Coração	160	26	5	14	5.9
I_9	Perna de ovelha/A	265	20	20	9	2.6
I_{10}	Mão de ovelha/A	300	18	25	9	2.3
I_{11}	Presunto/D	340	20	28	9	2.5
I_{12}	Porco/A	340	19	29	9	2.5
I_{13}	Porco/FE	355	19	30	9	2.4
I_{14}	Língua	205	18	14	7	2.5
I_{15}	Costela de vitela	185	23	9	9	2.7
I_{16}	Bluefish/C	135	22	4	25	0.6
I_{17}	Amêijoas/cru	70	11	1	82	6
I_{18}	Amêijoas/E	45	7	1	74	5.4
I_{19}	Siri/E	90	14	2	38	0.8
I_{20}	Haddock/F	135	16	5	15	0.5
I_{21}	Cavala/G	200	19	13	5	1
I_{22}	Cavala/E	155	16	9	157	1.8
I_{23}	Perca/F	195	16	1	14	1.3
I_{24}	Salmão/E	120	17	5	159	0.7
I_{25}	Sardinha/E	180	22	9	367	2.5
I_{26}	Atum/E	170	25	7	7	1.2
I_{27}	Camarão/E	110	23	1	98	2.6

Tabela 5.34: Dados de nutrientes em Carnes e Peixes por cada 3 onças (A=Assado, C=Cozido, D=Defumado, E=Enlatado, F=Frito, FE=Fervido, G=Grelhado).

estas duas partições é de 0.6141.

A figura 5.9 representa o dendrograma obtido pelo método de Ward de Classificação Hierárquica. A partição em duas classes, sugerida pelo critério de Ward é:

- Classe A: I_1 ; I_3 ; I_4 ; $I_{10} - I_{13}$;

	Classe A	Classe B
NHVG	$I_1 - I_4; I_9 - I_{14}; I_{21}$	$I_5 - I_8; I_{15} - I_{20}; I_{22} - I_{27}$
NHVC	$I_1 - I_4; I_9 - I_{14}; I_{21}$	$I_5 - I_8; I_{15} - I_{20}; I_{22} - I_{27}$
NHVMED	$I_1 - I_4; I_9 - I_{14}; I_{21}$	$I_5 - I_8; I_{15} - I_{20}; I_{22} - I_{27}$
NHMEAN	$I_1 - I_4; I_9 - I_{14}; I_{21}$	$I_5 - I_8; I_{15} - I_{20}; I_{22} - I_{27}$
K-MÉDIAS	$I_1 - I_4; I_9 - I_{13}$	$I_5 - I_8; I_{14} - I_{27}$

Tabela 5.35: Dados Nutrientes - Partição em 2 classes obtida pelos Métodos VL e *k*-Médias.

	Classe A	Classe B	Classe C
NHVG	$I_1 - I_4; I_9 - I_{14}$	$I_5 - I_8; I_{15} - I_{19}; I_{22}; I_{25}; I_{27}$	$I_{20}; I_{21}; I_{23}; I_{24}; I_{26}$
NHVC	$I_1 - I_4; I_9 - I_{14}$	$I_5; I_8; I_{15}; I_{17}; I_{18}; I_{25}; I_{27}$	$I_6; I_7; I_{16}; I_{19} - I_{24}; I_{26}$
NHVMED	$I_1 - I_4; I_9 - I_{14}$	$I_5; I_8; I_{15}; I_{17}; I_{18}; I_{25}; I_{27}$	$I_6; I_7; I_{16}; I_{19} - I_{24}; I_{26}$
NHMEAN	$I_1 - I_4; I_9 - I_{14}$	$I_5; I_8; I_{17}; I_{18}$	$I_6; I_7; I_{15}; I_{16}; I_{19} - I_{27}$
K-MÉDIAS	$I_1 - I_4; I_9 - I_{13}$	$I_{17}; I_{18}$	$I_5 - I_8; I_{14} - I_{16}; I_{19} - I_{27}$

Tabela 5.36: Dados Nutrientes - Partição em 3 classes obtida pelos Métodos VL e *k*-Médias.

- Classe B: I_2 ; $I_5 - I_9$; $I_{14} - I_{27}$.

Considerando a partição em três classes, obtém-se:

- Classe A: I_1 ; I_3 ; I_4 ; $I_{10} - I_{13}$;

- Classe B: I_{17} ; I_{18} ;

- Classe C: I_2 ; $I_5 - I_9$; $I_{14} - I_{16}$; $I_{19} - I_{27}$.

A partição mais próxima da partição sugerida pelo critério de Ward é a obtida pelo método das *k*-médias: ambas as partições são constituídas por uma classe que engloba as amêijoas (cru) e amêijoas (enlatado) que se caracterizam por apresentarem um baixo valor de gordura e energia e um elevado valor de ferro. O valor do IRM que mede a proximidade entre a partição obtida pelo critério de Ward e o método das *k*-médias é de 0.7378.

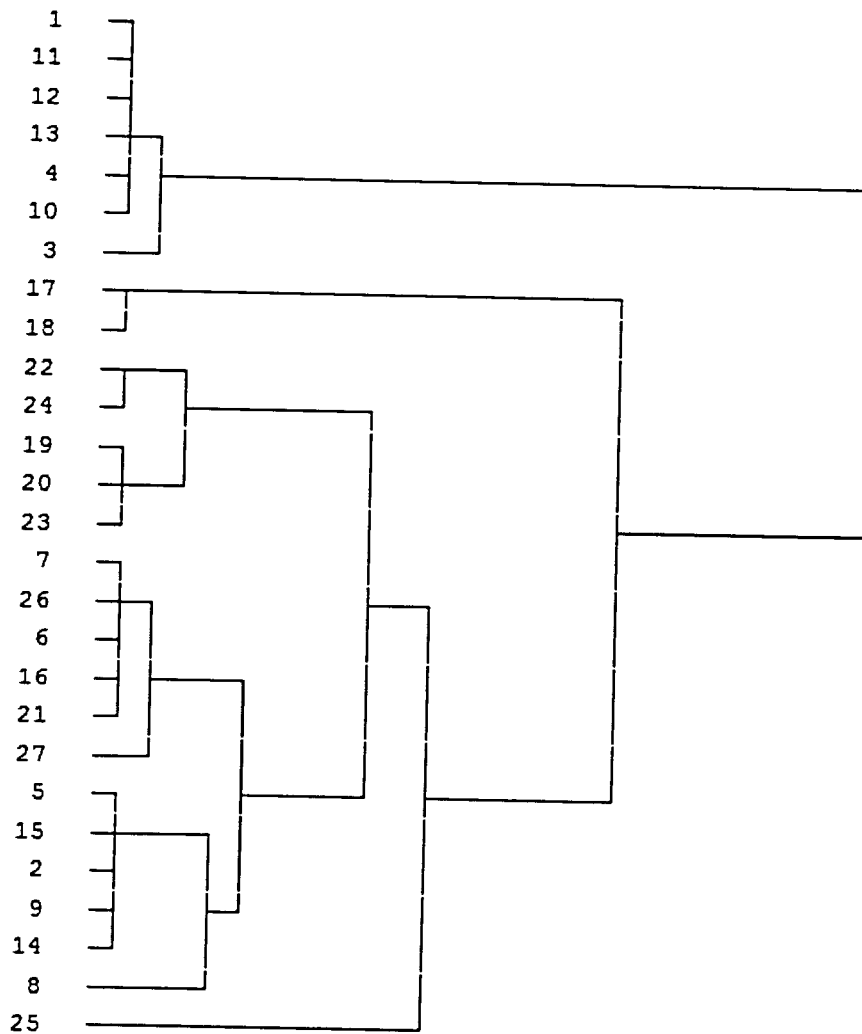


Figura 5.9: Dados Nutrientes - Classificação pelo Critério de Ward.

Capítulo 6

Estrutura do Programa

Com o intuito de estudar o comportamento dos métodos NHVC, NHVMED, NHVGM, NHMEAN e dos métodos directos associados foi elaborado um programa de classificação implementado em FORTRAN PowerStation versão 4.0.

O programa CLASSIFICAÇÃO constituído por várias subrotinas e funções, contempla a aplicação dos vários métodos referidos. Pretendendo comparar o desempenho dos novos métodos com os já existentes foram inseridas as subrotinas correspondentes a dois métodos de Classificação Não Hierárquica, o método NHMEAN e o método das k -Médias de MacQueen, disponíveis em Brito [9]. Paralelamente ao programa CLASSIFICAÇÃO foi implementado o programa PROG que tal como o primeiro, visa a classificação de um conjunto de dados segundo um dos métodos expostos, permitindo no entanto, a adaptação sistemática do número de classes ao longo do algoritmo (por fusão e/ou refinamento).

6.1 Unidade Principal do Programa

Nesta secção, para além de se fornecer informação adicional sobre as técnicas e algoritmos usados neste trabalho, pretende-se que o utilizador use o programa eficientemente e, se necessário, adapte o programa para futuras aplicações. O algoritmo é iniciado com um "diálogo interactivo" (programa-utilizador) a partir do qual o utilizador terá que definir determinados parâmetros de entre um conjunto de opções disponíveis. O programa verifica se cada opção feita pelo utilizador é uma resposta válida; se não fôr o caso, surge uma mensagem de erro e a questão é de novo colocada.

Os métodos de Classificação Não Hierárquica pressupõem ainda a definição de al-

guns parâmetros, prévia à classificação, tais como o número de classes pretendido e o número máximo de iterações permitido para convergência do método. Além disso, a definição do critério de escolha dos centros iniciais deve ser ainda efectuada.

6.1.1 Opções de Entrada

As opções de entrada do programa são:

- matriz de dados $X_{n \times p}$, n indivíduos $\times p$ variáveis, que poderá ser uma matriz de dados de escala real ou dados binários;
- matriz triangular inferior/superior de semelhanças/dissemelhanças, que será tratada como um vector de dimensão $\frac{n(n-1)}{2}$;
- função de comparação entre elementos, do tipo semelhança ou dissemelhança;
- método de afectação do tipo VL ou directo.

6.1.2 Opções de Inicialização dos Centros

De entre a panóplia de critérios de escolha de centros iniciais disponíveis na literatura, foram incluídos os seguintes:

1. Os centros são os k primeiros elementos da amostra.
2. Os centros são os elementos de ordem $\frac{j \times n}{k}$, $j = 1, \dots, k$.
Como já foi anteriormente referido, se estes valores não forem inteiros consideram-se os valores característicos.
3. Os centros são escolhidos aleatoriamente. Esta escolha é feita por um processo que gera números pseudo-aleatórios em $[0, 1]$, com posterior conversão para o conjunto $\{1, \dots, n\}$, pela função $x \mapsto y = INT[(n-1) * x] + 1$, $x \in [0, 1]$.
4. Os centros são fornecidos pelo utilizador.
5. Os centros são os elementos simultaneamente mais afastados na amostra e mais afastados entre si.
6. Critério de Lerman.
7. Critério *wise-choice*.

Note-se que os métodos considerados envolvem o cálculo de uma matriz de semelhanças entre elementos, permitindo assim, a aplicação dos critérios 5, 6 e 7.

6.1.3 Comparação de Partições: critério de estabilidade

O critério escolhido para convergência dos métodos de classificação estudados foi o da estabilidade da partição obtida, isto é, quando em duas partições sucessivas, as classes obtidas não sofrem qualquer alteração na sua constituição. Considera-se que um indivíduo permanece na mesma classe se se mantiver com pelo menos metade dos elementos que constituíam a classe à qual pertencia na partição anterior. Assim sendo, o número de transferências ocorridas de uma partição para a seguinte é dado por (cf. Brito [9])

$$\text{card } \{I_\ell \in E : A(I_\ell) = A(I_i) \wedge A'(I_\ell) \neq A'(I_i)\}$$

representando $A'(I_i)$ e $A(I_i)$ as classes que contêm o indivíduo I_i , respectivamente na partição actual e na partição anterior. Considera-se que o indivíduo I_i mudou de classe quando esse cardinal é superior a $\frac{\text{card } A(I_i)}{2}$.

6.1.4 Subrotinas

Subrotina Leitura

Esta subrotina contém a parte interactiva do programa, através da qual são introduzidos os parâmetros necessários.

Subrotinas LerX e LerP

Em qualquer uma destas subrotinas é feita a leitura dos dados. Perante uma matriz de dados, $X_{n \times p}$, é chamada a subrotina LerX. O registo de dados é feito de forma matricial se se tratarem de dados binários. No caso de se ter um quadro de dados métricos, cada elemento de uma linha do ficheiro é registado, obtendo-se uma única coluna com $n \times p$ elementos.

Se estiver disponível a matriz de semelhanças ou dissemelhanças entre elementos, o seu registo é feito também de modo sequencial, um elemento por linha do ficheiro, registando-se linha após linha da matriz em causa. A leitura é efectuada com a subrotina LerP.

Subrotinas ESTAT e CENTRED

Estas duas subrotinas permitem a centragem e redução de dados descritos por variáveis reais. Antes de qualquer alteração dos dados, é calculado o vector das médias e a matriz de variâncias e covariâncias das variáveis, com a subrotina ESTAT. Posteriormente, os dados são centrados e reduzidos.

Subrotinas SEMEL01 e SEMEL

Calculam a matriz de semelhanças VL entre indivíduos de uma amostra. Se os dados forem binários é chamada a subrotina SEMEL01, caso contrário, é usada a subrotina SEMEL. A fim de maximizar a memória disponível, é apenas armazenada a matriz triangular inferior.

Subrotinas NHMEAN, NHVCENTER, NHVMEDIAN e NHVGMEAN

Estas subrotinas contemplam a classificação, propriamente dita, de acordo com o critério de afectação definido. As subrotinas NHMEAN, NHVCENTER, NHVMEDIAN e NHVGMEAN efectuem a classificação pelo método NHMEAN, NIIVC, NIIVMED e NHVGM, respectivamente. As subrotinas apresentam uma estrutura similar, recorrendo, na etapa do cálculo da estatística amostral em causa e do valor da função de distribuição que lhe está associada, a diferentes cálculos numéricos, subrotinas e funções, como é óbvio. Assim, em NHVGMEAN, para o cálculo da função de distribuição qui-quadrado, foi introduzida a função FDQUI2, na qual são usadas as funções reais GAMMP e GAMMLN. Recorreu-se à subrotina MEDIAN para calcular a mediana de um vector de valores reais, não introduzindo qualquer alteração no vector. Em NIIVMEDIAN, para o cálculo do valor da função de distribuição no ponto obtido com a subrotina MEDIAN, foram introduzidas as funções reais FDBETA, FDMEDPAR, BETAI e BETACF. Todas estas subrotinas foram desenvolvidas no âmbito do trabalho Sousa [40].

Obtida uma partição, é dado lugar à avaliação global e local da partição obtida. Nas subrotinas MEDHET e MEDISO, estão disponíveis medidas de avaliação da homogeneidade e isolamento de classes, respectivamente, referidas no Capítulo 4.

A partição é comparada com uma partição referência (se existente) através da subrotina RANDMOD, a qual fornece o valor do Índice de Rand Modificado.

Subrotinas NHDMEAN, NHDCENTER, NHDMEDIAN e NHDGMEAN

Com estas subrotinas classifica-se o conjunto de dados pelos Métodos Directos (Secção 3.2). As subrotinas NHDMEAN, NHDCENTER, NHDMEDIAN e NHDGMEAN efectuem a classificação pelo método NHDMEAN, NHDC, NHDMED e NHDGM, respectivamente.

Subrotinas NCLMEAN, NCLC, NCLMED e NCLGM

O objectivo destas rotinas é o de testar o número de classes escolhido inicialmente pelo analista, efectuando eventualmente a fusão de classes. As subrotinas NCLMEAN, NCLC, NCLMED e NCLGM, estão inseridas na subrotinas NHMEAN, NHVCENTER, NHVMEDIAN e NHVGMEAN, respectivamente.

Subrotinas CENTROMEAN, CENTROC, CENTROMED e CENTROGM

Após o processo de afectação, se o critério de convergência não é atingido, prossegue-se com uma nova iteração do algoritmo. Neste caso, é necessário o cálculo dos centros das classes obtidas na etapa anterior, o que é possível com as subrotinas CENTROMEAN, CENTROC, CENTROMED e CENTROGM, nos critérios NHMEAN, NHVC, NHVMED e NHVGM, respectivamente.

Capítulo 7

Notas Finais

Desenvolvimentos Futuros

No trabalho desenvolvido propuseram-se e estudaram-se novos métodos no âmbito da metodologia VL, no quadro da Classificação Não Hierárquica. Estes métodos recorrem à função de distribuição de estatísticas de tendência central, das semelhanças/dissimelhanças entre pares de elementos do conjunto a classificar, como critério de afectação de um elemento a uma classe. Os métodos propostos foram os métodos NHVGM, NIIVMED e NHVC, que usam as estatísticas amostrais média geométrica, mediana e centro, respectivamente. Tendo em conta a estatística usada, foram sugeridas adaptações ao critério de inicialização *wise-choice*.

Como conclusão geral dos resultados obtidos, pode referir-se que o desempenho dos métodos propostos depende fortemente do tipo de estrutura do conjunto a classificar. Analogamente, comparando os resultados obtidos com os métodos denominados "probabilísticos" e com os métodos denominados "directos", concluiu-se que a existência de vantagem no recurso à função de distribuição na função de afectação dos elementos às classes depende igualmente da estrutura dos dados.

A aplicação dos métodos VL a dados simulados permitiu concluir que, dos novos métodos propostos, o método NIIVMED foi o que, em geral, melhor desempenho obteve, produzindo resultados bastante satisfatórios. Os métodos NHVC e NHVGM nem sempre foram capazes de recuperar a estrutura subjacente aos conjuntos analisados, facto resultante do "efeito de massificação" já verificado em métodos de Classificação Hierárquica. Foi ainda possível confirmar que o grau de separação das classes

influencia a qualidade da partição obtida.

O uso do critério de inicialização *wise-choice*, com as adaptações sugeridas (Secção 3.3) foi motivador, uma vez que para os três novos métodos o número de iterações necessárias para atingir a convergência é reduzido. No sentido de validar os critérios de inicialização propostos, um estudo sobre os mesmos conjuntos de dados simulados poderá ser levado a cabo, tendo em conta o mesmo critério de inicialização sem as referidas adaptações, ou recorrendo a outros critérios de inicialização.

Como perspectiva de desenvolvimento futuro impõe-se um estudo detalhado do efeito de massificação nos métodos VL, em particular nos métodos NIIVC e NIIVGM. A análise do desempenho dos métodos VL incidiu sobre conjuntos de dados simulados com distribuição binormal tendo em conta diferentes formas de classes e graus de separação. Não tendo sido este estudo exaustivo, outros conjuntos com diferentes distribuições poderão ser estudados, variando ainda o cardinal e a forma das classes, assim como o número de variáveis.

Outro ponto de desenvolvimento a considerar prende-se com o uso de estatísticas alternativas que poderão constituir novas opções para a definição da função de afectação dos indivíduos às classes.

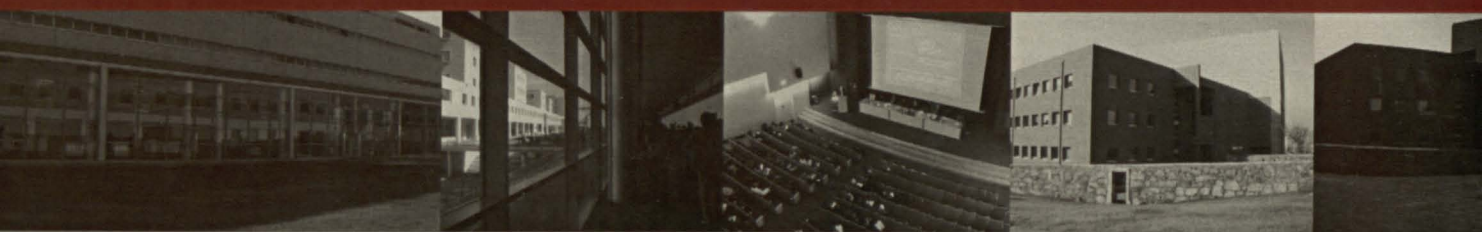
Bibliografia

- [1] Anderberg, M. R. (1973) - *Cluster Analysis for Applications*, Academic Press, New York.
- [2] Arabie, P. ; Hubert, L. J. ; Soete, G. (1996) - *Clustering and Classification*, World Scientific, Singapore.
- [3] Bacelar-Nicolau, H. (1972) - *Analyse d'un Algorithme de Classification Automatique - Thèse de 3ème cycle*, ISUP, Université Paris VI.
- [4] Bacelar-Nicolau, H. ; Nicolau, F.C. (1994) - *Hierarchical Clustering Models and Robustness*, *Actes des XXVI Journées de Statistique*, A.S.U., Neuchâtel, p. 96-99.
- [5] Ball, G.H. ; Hall, D.J. (1967) - *A Clustering Technique for Summarizing Multivariate Data*, *Behavioral Science*, Vol. 12.
- [6] Beale (1969) - *Euclidean Cluster Analysis*, *Bull. I.S.I.*, 43, 2, p. 92-94.
- [7] Benzécri, J. P. (1973) - *L'Analyse des Données, Tome 1: La Taxonomie*, Dunod, Paris.
- [8] Bock, H. - H. (1974) - *Automatische Klassifikation*, Vandenhoeck & Ruprecht, Goettingen, cap. 17.
- [9] Brito, P. (1988) - *Métodos Não Hierárquicos de Classificação - O Método NIIMEAN*, Dissertação de Mestrado, Faculdade de Ciências, Universidade de Lisboa.
- [10] Celeux, G. ; Diday, E. ; Govaert, G. ; Lechevallier, Y. ; Ralambondrainy, H. (1989) - *Classification Automatique des Données - Environnement Statistique et Informatique*, Dunod, Paris.

- [11] Diday, E. (1972) - *Nouveaux Concepts et Nouvelles Méthodes en Classification Automatique*- Thèse d' état, Univ. Paris VI.
- [12] Diday, E. ; Simon, J. J. (1976) - *Clustering Analysis*, K. S. Fu (Ed.), Digital Pattern Recognition, Springer, Heidelberg, p. 47-94.
- [13] Fisher, L. ; Van Ness, J. (1971) - *Admissible Clustering Procedures*, Biometrika 58, p. 91-104.
- [14] Fisher, W.D. (1958) - *On Grouping for Maximum Homogeneity*, J.A.S.A., Vol. 53, p. 789-798.
- [15] Fisz, M. (1963) - *Probability Theory and Mathematical Statistics*, 3rd Edition, Wiley Publications in Statistics.
- [16] Forgy, E.W. (1965) - *Cluster Analysis of Multivariate Data: Efficiency vs. Interpretability of Classifications*, in WNAR Meeting, University of California, Riverside, June 22-23, 1965; abstract in Biometrics, Vol. 21, 3, p. 768.
- [17] Gordon, A. D. (1999) - *Classification*, 2nd Edition, Chapman & Hall, U.S.A..
- [18] Guénoche, A. ; Garreta, H. (2002) *Representation and Evaluation of Partitions*, in Classification, Clustering and Data Analysis; Jajuga, K. ; Sokolowski, A. ; Bock, H.-H. (Eds), p. 131-138.
- [19] Hansen, P. ; Jaumard, B. (1997) - *Cluster Analysis and Mathematical Programming*, Mathematical Programming, 79, p. 191-215.
- [20] Hartigan, J. (1976) - *Clustering Algorithms*, John Wiley and Sons, London.
- [21] Hubert, L. ; Arabie, P. (1985) - *Comparing Partitions*, Journal of Classification, 2, p. 193-218.
- [22] Jancey, R.C. (1966) - *Multidimensional Group Analysis*, Austr. J. Bot., Vol. 14, N° 1, p. 127-130.
- [23] Jardine, N. ; Sibson, R. (1967) - *Mathematical Taxonomy*, John Wiley and Sons, London.
- [24] Kaufman, L. ; Rousseeuw, P. J. (1990) - *Finding Groups in Data - An introduction to Cluster Analysis*, Wiley Series in Probability and Mathematical Statistics.

- [25] Kendall & Stuart (1976) - *The Advanced Theory of Statistics*, Vol. 3, 3ª ed.
- [26] Lerman, I. C. (1970) - *Sur l'Analyse des Données Préalable à une Classification Automatique*, *Mathématiques et Sciences Humaines*, vol. 32, p. 5-15, Paris.
- [27] Lerman, I. C. (1981) - *Classification et Analyse Ordinale des Données*, Dunod, Paris.
- [28] MacQueen, J. (1967) - *Some Methods for Classification and Analysis of Multivariate Observations*, Proc. of the 5th Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, p. 281-297.
- [29] Nicolau, F. C. (1980) - *Critérios de Análise Classificatória Hierárquica Baseados na Função Distribuição*, Tese de Doutoramento, Faculdade de Ciências, Universidade de Lisboa.
- [30] Nicolau, F. C. (1981)- *Métodos Não Hierárquicos de Classificação - O Método das K-Médias*, Prova Complementar de Doutoramento, Faculdade de Ciências, Universidade de Lisboa.
- [31] Nicolau, F. C. (1985) - *Analysis of a Non Hierarchic Clustering Method Based on VL-Similarity*, Centro de Estatística e Aplicações/ I.N.I.C., Nota Nº 18/85.
- [32] Nicolau, F. C. ; Bacelar-Nicolau, H. (1981) - *Novos Aspectos da Análise Classificatória*, *Actas do II Colóquio de Estatística e Investigação Operacional*, Fundão.
- [33] Nicolau, F. C. ; Brito, P. (1989) - *Improvements in NIIMEAN Method*, in *Data Analysis, Learning Symbolic and Numeric Knowledge*, Diday, E. (Ed.), Nova Science Publishers Inc., New York, p. 109-116.
- [34] Ok-Sakun, Y. (1975) - *Analyse Factorielle Typologique et Lissage Typologique*, Thèse de 3ème cycle, Univ. Paris VI.
- [35] Régnier, S. (1965) - *Sur Quelques Aspects Mathématiques des Problèmes de la Classification Automatique*, I.C.C. Bulletin, Vol. 4.
- [36] Roux, M. (1985) - *Algorithmes de Classification*, Masson.
- [37] Rubinstein, R. Y. (1981) - *Simulation and the Monte Carlo Method*, Wiley Series in Probability and Mathematical Statistics.

- [38] Saporta, G. ; Youness, G. (2004) - *Une Méthodologie Pour la Comparaison de Partitions*, Rev. Statistique Appliquée, 2004, LII(1), p. 97-120.
- [39] Schroeder, A. (1976) - *Analyse d'un Mélange de Distributions de Probabilité de Même Type*, Rev. Statistique Appliquée, 24, 1, p. 39-62.
- [40] Sousa, F. C. (2000) - *Novas Metodologias e Validação em Classificação Hierárquica Ascendente*, Tese de Doutoramento, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa.
- [41] Van Ness, J. W. (1999) - *Recent Results in Clustering Admissibility in Applied Stochastic Models and Data Analysis - Quantitative Methods in Business and Industry Society*, Proc. of the International Symposium on Applied Stochastic Models and Data Analysis (ASMDA - 99), Lisbon, Portugal, Bacelar Nicolau, H.; Nicolau, F.C.; Janssen, J. (Eds.), Instituto Nacional de Estatística, Lisboa, Portugal, p. 19-29.



FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

BIBLIOTECA



0000072776