Sílvia Maria de Sousa Amorim

A escolha do número de classes no Método de Classificação das *k*-Médias

Departamento de Matemática Aplicada
Faculdade de Ciências da Universidade do Porto
Maio/2001

Sílvia Maria de Sousa Amorim

A escolha do número de classes no Método de Classificação das *k*-Médias

Departamento de Matemática Aplicada
Faculdade de Ciências da Universidade do Porto
Maio/2001

Sílvia Maria de Sousa Amorim

A escolha do número de classes no Método de Classificação das *k*-Médias

Tese submetida à Faculdade de Ciências da Universidade do Porto
Para a obtenção do grau de Mestre em Estatística

Departamento de Matemática Aplicada
Faculdade de Ciências da Universidade do Porto
Maio/2001

Dissertação orientada por

Joaquim Costa

Professor Auxiliar do
Departamento de Matemática Aplicada da
Faculdade de Ciências da Universidade do Porto

e por Luís Torgo

Professor Auxiliar do
Grupo de Matemática e Informática da
Faculdade de Economia da Universidade do Porto

Aos meus pais

I. Agradecimentos

No momento em concluí esta tese, não queria deixar de agradecer aos professores Doutor Joaquim Costa e Doutor Luís Torgo pelas suas orientações e pelo seu permanente estímulo, o que tornou possível este estudo e à Drª. Helena Brás pelo seu interesse, simpatia e colaboração na parte prática deste trabalho.

Agradeço também aos meus colegas de Mestrado com quem partilhei experiências muito gratificantes, quer em termos profissionais, quer em termos pessoais e em especial aos meus amigos que tão bem souberam incentivar: um telefonema, uma palavra dita na ocasião certa, um bater de ombro... E a todos aqueles que comigo colaboraram, pela sua infinita disponibilidade e dedicação.

Por último, um agradecimento especial à minha família, pela sua compreensão e apoio incondicional, essencialmente, nos momentos mais dificeis.

II. Resumo e Palavras Chave

O método de agrupamento das k-médias é uma das técnicas disponíveis para, dado um conjunto de observações, obter uma partição dessas observações em k grupos. Este método funciona através de um processo de minimização de uma função que calcula um valor que resulta da soma das distâncias entre cada observação e o centro do grupo a que estão actualmente alocadas. Com base neste valor o método tenta deslocar as observações de um grupo para o outro de modo a minimizar o valor desta função. Um dos problemas levantados por este método de agrupamento é o de descobrir o valor óptimo para k (isto é, quantos grupos devem ser formados).

Neste trabalho iremos estudar o método de agrupamento das k-médias, bem como os diferentes problemas levantados por esta metodologia (o valor do k; as medidas de distância entre as observações; etc.)

Palavras-chave: Análise Classificatória, k-médias, critério de Lerman, escolha de k

III. Abstract and Keywords

The k-means algorithm for clustering is one of the available techniques to obtain a partition of a given set of observations in k groups. This method works by minimizing a function which estimates a value resulting of the sum of distances between each observation and the centre of the group to which it is allocated. Based on this value, the method tries to move the observations from one group to another in order to minimize the value of the function. One of the problems raised by this method is finding out the appropriate value for k (i. e. how many groups should be formed).

In this study we are going to analyse the k-means algorithm for clustering, as well as the different questions raised by this methodology (the value of k, the measures of the distances between observations, etc.).

Keywords: Clustering, k-means, Lerman index, choice of k

IV. Índice

1. Introdução

Uma das actividades mais antigas e comuns do Homem consiste em classificar objectos por classes.

Hoje em dia, a classificação está dividida em dois tipos: Análise Discriminante (ou Classificação Supervisionada) e Análise Classificatória (ou Classificação não supervisionada).

O objectivo é classificar observações, isto é, reconhece-las como membros de uma classe. No entanto, são de natureza diferentes; em Análise Discriminante, as classes dos indivíduos são conhecidas, queremos construir uma regra de classificação que nos permita prever a classe de indivíduos futuros; enquanto que, em Análise Classificatória não se conhecem as classes.

Assim, em Análise Classificatória, dado um conjunto de dados o objectivo é encontrar uma estrutura de classes que se ajuste a estas observações. O problema é muitas vezes exposto como o de encontrar uma "estrutura natural" de classes que se ajuste a estas observações. Mais concretamente, o objectivo é dispor as observações em grupos de tal forma que o seu grau de "associação natural" seja grande entre os elementos do mesmo grupo e pequeno para elementos de grupos diferentes. É este tipo de classificação que iremos abordar neste trabalho.

Na prática, só conseguimos saber se uma estrutura é boa ou má quando a conhecemos. Uma solução seria enumerar todas as possíveis estruturas e escolher a mais indicada. Mas isto torna-se computacionalmente impossível de realizar a não ser para conjuntos de dados muito pequenos.

O número de maneiras de distribuir n observações em m grupos é um número de Stirling do 2º tipo; ou seja,

$$S_n^{(m)} = \frac{1}{m!}\sum_{k=0}^{k=m}(-1)^{m-k}\binom{m}{k}k^n \quad \text{(Abramowitz e Stegun, 1968).}$$

Assim, por exemplo, para distribuir 25 observações em 5 classes o número de possibilidades é:

$$S_{25}^{(5)} = 2436684974110751.$$

O problema pode ser ainda agravado pelo facto de o número de grupos não ser conhecido, passando assim o número de grupos a ser uma soma de números de Stirling. No caso de 25 observações é

$$\sum_{j=1}^{25} S_{25}^{(j)} > 4 \times 10^8 !$$

Houve realmente, uma necessidade, de desenvolver métodos que nos permitissem classificar um dado conjunto de objectos.

2. Análise Classificatória

O principal objectivo da Análise Classificatória (ou Classificação não supervisionada) é identificar estruturas ou classes presentes no conjunto de dados. Ou seja:

Dado um conjunto E de n objectos descritos por p variáveis, pretende-se agrupá-los em k classes; de forma a que objectos de um mesmo grupo sejam mais semelhantes do que objectos de grupos distintos.

Existem muitas formas de o fazer; pretende-se geralmente que as classes fiquem bem separadas.

2.1. Análise Classificatória Paramétrica

Seja $X_1, X_2, \ldots, X_n$ a amostra a classificar.

$X_i^T = (x_{i1}, x_{i2}, \ldots, x_{ip})$

Pretende-se agrupar as observações em k classes.

Seja k_i a classe atribuída a X_i; portanto, $k_i \in \{1,2,\ldots,k\}$.

Seja $\Omega = (k_1, k_2, \ldots, k_n)^T$ o vector das classificações e $X^* = \left(X_1^T, X_2^T, \ldots, X_n^T\right)^T$ o vector das configurações.

Para obter a classificação temos que optimizar um dado critério J:

$$J = J(\Omega, X^*).$$

Por definição, a melhor classificação, Ω_0, é tal que

$$J(\Omega_0, X^*) = \max_{\Omega} \text{ou} \min_{\Omega} J(\Omega, X^*).$$

- Matriz de dispersão dentro das classes:

$$S_W = \sum_{i=1}^{k} \pi_i E\{(X-\mu_i)(X-\mu_i)^T / C_i\} = \sum_{i=1}^{k} \pi_i \Sigma_i .$$

Onde π_i é a probabilidade da classe *i* e μ_i é o valor médio da classe *i*. Σ_i é a matriz de variâncias-covariâncias para a classe *i* e C_i a classe *i*.

S_W mostra a dispersão das amostras em torno dos respectivos centros.

- Matriz de dispersão entre as classes:

$$S_B = \sum_{i=1}^{k} \pi_i (\mu_i - \mu_0)(\mu_i - \mu_0)^T ,$$

onde μ_0 é o vector médio total.

S_B mede a dispersão dos vários vectores médios em torno do vector médio total.

- Matriz de dispersão mistura:

Esta matriz é a matriz de covariância de todas as amostras independentemente da classe:

$$S_m = E\{(X-\mu_0)(X-\mu_0)^T\} = S_W + S_B .$$

Utilizamos estas matrizes para medir a separabilidade das classes; para isso, é comum utilizar critérios do tipo:

a) $J = tr\left(S_2^{-1}S_1\right)$;

b) $J = \ln\left|S_2^{-1}S_1\right| = \ln|S_1| - \ln|S_2|$;

c) $J = trS_1 - \mu(trS_2 - c)$;

d) $J = \dfrac{trS_1}{trS_2}$;

onde S_1 e $S_2 \in \{S_B, S_W, S_m\}$.

Para obter a classificação, usamos um algoritmo iterativo.

Para isso, e sem perda de generalidade, vamos supor que se pretende minimizar o critério J.

2.2. Algoritmo

1) Escolha uma classificação inicial

$$\Omega(0) = (k_1(0), k_2(0), \ldots, k_n(0))^T.$$

2) Seja $\Omega(l)$ classificação na iteração l

$$\Omega(l) = (k_1(l), k_2(l), \ldots, k_n(l))^T.$$

Seja $\Delta J(i,j,l)$ a variação do critério se a observação i mudar da sua classe actual, $k_i(l)$, para a classe j:

$$\Delta J(i,j,l) = J((k_1(l),k_2(l),...,k_{i-1}(l),j,k_{i+1}(l),...,k_n(l)),X^*) - J(\Omega(l),X^*),$$

onde $i = 1,2,...,n$ e $j = 1,2,...,k$.

3) Para cada $i = 1,2,...,n$ seja

$$\Delta J(i,t,l) = \min_j \Delta J(i,j,l)$$

então, mudar a observação X_i para a classe t.

Obtém-se assim $\Omega(l+1)$.

4) Se $\Omega(l+1) \neq \Omega(l)$, voltar a 2), senão parar.

Notas:

- No passo 3), as amostras podem ser reclassificadas todas ao mesmo tempo ou não. Os resultados são ligeiramente diferentes.

- Sob certas condições o algoritmo converge.

- Mesmo que haja convergência, não temos a certeza de termos obtido o menor valor de J.

Apesar disto, o algoritmo é bastante eficiente.

2.3. Determinação do número de classes

Na prática, em geral não conhecemos k. Uma solução seria, para cada $k \geq 2$ encontrar a melhor classificação; seja $J^*(k)$ o valor do critério correspondente. No entanto, com a maioria dos critérios, $J^*(k)$ decresce sempre. Por exemplo, $J = tr\left(S_m^{-1} S_W\right)$ é nulo quando $k = n$.

Como podemos controlar k?

Uma forma de o fazer consiste em escolher um k inicial e depois unir ou separar classes.

Se duas classes forem muito semelhantes, elas podem ser unidas; para medir a semelhança entre classes, podemos usar a distância euclidiana entre os vectores médios ou a distância de Battacharyya:

$$d_B(C_1, C_2) = \frac{1}{8}(\mu_2 - \mu_1)^T \left(\frac{\Sigma_1 + \Sigma_2}{2}\right)^{-1} (\mu_2 - \mu_1) + \frac{1}{2} \ln \frac{\left|\frac{\Sigma_1 + \Sigma_2}{2}\right|}{\sqrt{|\Sigma_1||\Sigma_2|}}.$$

Também quando uma classe contém poucos elementos, ela pode ser unida à classe mais semelhante.

Para dividir uma classe em duas ou mais classes, é mais difícil; geralmente, considera-se a divisão quando uma classe tem muitos elementos, é multimodal ou, tem uma grande variância numa dada direcção. Para efectuar a divisão, pode-se aplicar Análise Classificatória somente aos elementos dessa classe.

Vamos considerar agora um critério particular:

$$J = tr\left(S_m^{-1} S_B\right) \text{ ou } J = tr\left(S_m^{-1} S_W\right).$$

Estes critérios são equivalentes, pois

$$J = tr\left(S_m^{-1}S_B\right) = tr\left(S_m^{-1}\left(S_m - S_W\right)\right) = n - tr\left(S_m^{-1}S_W\right).$$

Logo, maximizar $tr\left(S_m^{-1}S_B\right)$ é equivalente a minimizar $tr\left(S_m^{-1}S_W\right)$.

Vamos também assumir que μ_0 é o vector média total e S_m a matriz identidade. Se os dados não verificarem isto, podemos sempre aplicar uma transformação linear:

$$Y = \Lambda^{-\frac{1}{2}}\Phi^T X \text{ (Fukunaga,1990).}$$

Onde Λ é a matriz diagonal dos valores próprios de S_m e Φ é a matriz dos vectores próprios de norma 1. Esta transformação linear faz com que a matriz de covariâncias de Y seja a identidade.

Utilizando a amostra $X_1, X_2, \ldots X_n$ (ou $Y_1, Y_2, \ldots, Y_n$ se for caso disso) vamos estimar as matrizes S_w e S_m.

- $\hat{S}_W = \frac{1}{n}\sum_{r=1}^{k}\sum_{j=1}^{n_r}\left(X_j^r - \mu_r\right)\left(X_j^r - \mu_r\right)^T$ e $\hat{S}_m = \frac{1}{n}\sum_{j=1}^{n}\left(X_j - \hat{\mu}_0\right)\left(X_j - \hat{\mu}_0\right)^T$.

$$J = tr\left(\hat{S}_m^{-1}\hat{S}_W\right) = tr\left(\hat{S}_W\right) = \frac{1}{n}\sum_{r=1}^{k}\sum_{j=1}^{n_r}\left(X_j^r - \mu_r\right)^T\left(X_j^r - \mu_r\right) =$$

$$= \frac{1}{n}\sum_{r=1}^{k}\sum_{j=1}^{n_r}\left\|X_j^r - \mu_r\right\|^2.$$

Se mudarmos X_i da sua classe, k_i, para a classe j, então

$$\Delta J(i,j,l) = \frac{1}{n}\left\{\|X_i - \mu_j\|^2 - \|X_i - \mu_{k_i}\|^2\right\}.$$

Minimizar $\Delta J(i,j,l)$ é equivalente a minimizar o primeiro termo, pois o segundo termo é constante.

Se $\|X_i - \mu_t(l)\| = \min_j \|X_i - \mu_j(l)\|$, então mudamos a observação X_i para a classe t.

Em resumo:

Algoritmo

1. Escolhe-se uma classificação inicial $\Omega(0)$ e calcula-se $\mu_1(0), \mu_2(0), \ldots, \mu_k(0)$. Ou então, escolhe-se directamente os centros iniciais.

2. Na iteração número $l + 1$, depois de calcular $\mu_1(l), \mu_2(l), \ldots, \mu_k(l)$, para cada $i = 1,2,\ldots,n$, reclassifica-se X_i na classe cuja média está mais próxima e logo de seguida, reactualiza-se o centro das duas classes.

3. Se não houver mudanças, para-se; senão voltar a 2.

Este é precisamente o **Algoritmo de Agrupamento das *k*-médias de MacQueen (1967).**

2.4. Propriedades do Algoritmo de Agrupamento das *k*-médias de MacQueen (1967)

- É evidente que as classes são divididas por hiperplanos, pois só as médias é que contam; as matrizes de covariância não interferem.

- O número de classes tem de ser pré-determinado.

- A classificação inicial é aleatória mas não causa instabilidade. De forma equivalente podemos escolher aleatoriamente os $\mu_i(0)$.

- No algoritmo de MacQueen (1967), os objectos são reclassificados um de cada vez; de cada vez que se reclassifica um objecto, os respectivos centros são actualizados. No algoritmo de Forgy (1965) temos que todos os objectos são reclassificados simultaneamente; só no fim de cada iteração é que os centros são recalculados.

- O custo computacional do algoritmo das *k*-médias é $O(Tkn)$ onde T é o número de iterações e n é o número de objectos do conjunto de dados (Anderberg, 1973). Por isso, este algoritmo é eficiente para processar grandes conjuntos de dados.

- O algoritmo das *k*-médias divide o espaço IR^p em regiões, que podem mais tarde servir para afectar novos indivíduos. A maior parte dos algoritmos de Análise Classificatória não possui esta propriedade predictiva.

- Tradicionalmente, o método das *k*-médias toma para centros iniciais os primeiros *k* indivíduos: pode-se, no entanto, usar outro critério.

- O algoritmo das *k*-médias apenas trabalha com dados numéricos; no entanto, Ralambondrainy (1995) apresentou uma aproximação do algoritmo das *k*-médias para dados qualitativos (Huang, 1998).

Existem algumas variantes deste algoritmo as quais diferem na selecção dos centros iniciais, no cálculo das dissemelhanças e nas estratégias para calcular os centros dos grupos (Anderberg, 1973; Bobrowski e Bezdek, 1991). Uma das variantes sofisticadas do algoritmo das *k*-médias inclui o algoritmo ISODATA (Ball e Hall, 1965), o método NHMEAN (Nicolau e Brito, 1989) e os algoritmos das *k*-médias difuso (Ruspini, 1969, 1973; Bezdek, 1981).

2.5. O algoritmo ISODATA

O algoritmo ISODATA foi desenvolvido, durante vários anos, no Stanford Research Institute.

Ball e Hall (1965) apresentaram uma completa descrição deste método e ilustraram-no detalhadamente com um exemplo de duas dimensões.

Este algoritmo tem sido sujeito a uma grande investigação e, por isso, existem várias versões dele.

A versão que vamos apresentar é a que maiores semelhanças tem com o algoritmo original mas, difere em alguns detalhes. O método consiste no seguinte:

0. Os valores que se seguem devem ser escolhidos previamente.

1. Escolhe-se os centros dos grupos. Fixa-se cada um dos objectos ao grupo de cujo centro está mais próximo.

2. Recalculam-se os centros dos grupos. Repete-se este procedimento e calculam-se os centros até haver convergência ou até o número desses ciclos atingirem um valor pré-fixado NPARTS.

3. Ignorar todos os grupos que contêm menos que THETAN elementos.

4. De seguida, realiza-se uma agregação ou uma separação de acordo com as seguintes regras:

Processo de fusão: para a partição presente numa dada etapa determina-se o par de centros mais próximos. Se a distância entre eles for inferior a C decide-se fundir as respectivas classes numa só classe, cujo centro é depois calculado. Este procedimento é repetido até que todos os centros distem entre si de pelo menos C. Assim, em cada etapa, o número de classes é eventualmente reduzido.

Analogamente ao que se passa no algoritmo de agrupamento das k-médias, o processo termina, quando depois de uma iteração não houver nenhuma alteração; este método é bastante eficiente.

O número de grupos com que se termina o processo é uma incógnita.

É preciso escolher os valores de C e R; em geral, toma-se $C \leq R$.

Um dos inconvenientes levantados pelo algoritmo de MacQueen é a necessidade de saber *a priori* o número de classes existentes na estrutura de dados, informação esta que nem sempre está disponível; é sobre esta questão que nos vamos centrar a partir de agora neste trabalho.

3. Partição óptima

Vamos agora descrever um método que nos permite encontrar uma partição óptima para um dado conjunto de objectos a classificar, desenvolvido por Lerman (1970, 1981).

Suponhamos que $E = \{X_1, X_2, ..., X_n\}$ é um conjunto de cardinal n. Seja $F = P_2(E) = \{\{X,Y\} : X,Y \in E, X \neq Y\}$ o conjunto dos pares de objectos distintos de E.

3.1. Comparação entre elementos

A comparação entre elementos é feita através, de um índice de comparação entre pares de elementos do conjunto E a classificar:

$$\gamma : E \times E \to IR .$$

Existem dois tipos de índices de comparação:

Índice de semelhança: grandes valores do índice representam elevada semelhança entre os elementos;

Índice de dissemelhança: grandes valores do índice representam afastamento entre os indivíduos.

Definição 3.1.:

Um índice de dissemelhança é uma função de comparação positiva $d : E \times E \to IR^+$, tal que:

$$d(X,X) = 0 ;$$

$$d(X,Y) = d(Y,X), \forall X,Y \in E.$$

Definição 3.2.:

Uma distância é um índice de dissemelhança, tal que:

$$d(X,Y) = 0 \Rightarrow X = Y;$$
$$d(X,Y) \leq d(X,Z) + d(Z,Y), \forall X,Y,Z \in E.$$

As noções correspondentes em termos de semelhança são:

Definição 3.3.:

Um índice de semelhança é uma função de comparação positiva $s : E \times E \rightarrow IR^+$, tal que:

$$s(X,X) = s_{max}, \forall X \in E;$$
$$s(X,Y) = s(Y,X), \forall X,Y \in E.$$

Definição 3.4.:

Proximidade é um índice de semelhança, tal que:

$$s(X,Y) = s_{max} \Rightarrow X = Y;$$
$$s_{max} + s(X,Y) \geq s(X,Z) + s(Z,Y), \forall X,Y,Z \in E.$$

O nosso objectivo é comparar duas partições sobre o conjunto E. Para isso, vamos começar por introduzir algumas notações.

Seja I um índice de semelhança entre os elementos de E (podia ser um índice de dissemelhança). Vamos considerar o caso em que o índice I induz uma relação de ordem total e estrita sobre o conjunto F:

$$p < q \Leftrightarrow I(p) < I(q), \forall p,q \in F . \quad \textbf{(3.1.)}$$

Designaremos esta relação por $W = W(E)$.

A relação de ordem (3.1.) pode ser vista como um subconjunto do produto cartesiano $F \times F$:

$$gr(W) = \{(p,q) : p,q \in F \text{ e } p < q \text{ para a relação de ordem } W\}$$

Tem-se portanto, que o índice de semelhança entre os elementos do conjunto que se pretendem classificar pode ser representado por um subconjunto de $F \times F$.

Consideremos agora uma partição π do conjunto E em k classes. Seja $t = (n_1, n_2, ..., n_k)$ o tipo dessa partição, onde n_i é o número de elementos da classe i.

Vamos agora definir um índice de proximidade ou de associação entre uma partição π e a relação de ordem W.

Seja $\pi = (E_1, E_2, ..., E_k)$, tal que $E_i \cap E_j = \{\}$ e $\bigcup_{i=1}^{k} E_i = E$.

Seja $R(\pi)$ o conjunto dos pares de objectos que pertencem a uma mesma classe da partição π:

$$R(\pi) = \{p = \{X,Y\} : \exists j \in \{1,...,k\} : X \in E_j ; Y \in E_j\} = \bigcup_{j=1}^{k} P_2(E_j).$$

É claro que $R(\pi)$ é um subconjunto de F. Seja $S(\pi)$ o conjunto complementar de $R(\pi)$ em F; $S(\pi)$ é o conjunto dos pares de objectos separados pela partição π:

$$S(\pi)=\{p=\{X,Y\}:\exists j\neq h: X\in E_j \text{ e } Y\in E_h\}=\bigcup_{1\leq j<h\leq k} E_j\times E_h .$$

Utilizando as notações assumidas, temos que a partição π pode ser representada por um subconjunto de $F\times F$:

$$S(\pi)\times R(\pi).$$

Assim, tanto a relação de ordem W como a partição π podem ser representadas por subconjuntos $F\times F$, o que facilita a sua comparação.

Com o objectivo de comparar a ordem W com a partição π, Lerman (1981, 1983) começa por definir um índice "bruto" de proximidade entre W e π:

$$s(W,\pi)=card[gr(W)\cap(S(\pi)\times R(\pi))]. \quad \textbf{(3.2.)}$$

De seguida, procura normalizar este índice.

Seja $P(n; t)$ o conjunto de todas as partições de E, que são do mesmo tipo que t de π.

Vamos agora introduzir a noção de parte aleatória de um conjunto associado a uma parte D de cardinal m de E. Lerman (1981, 1983) refere que existem três modelos aleatórios possíveis e designa-os por N_1, N_2 e N_3, respectivamente.

N_1: Para este modelo X é um elemento aleatório do conjunto das partes de E de cardinal m, $P_m(E)$. Existem $\binom{n}{m}$ partes nessas condições e a probabilidade é uniformemente repartida sobre $P_m(E)$.

N_2: Se considerarmos o conjunto das partes de E, organizado por inclusão, onde cada nível é formado pelos conjuntos com o mesmo cardinal; o modelo N_1 corresponde a transportar toda a probabilidade sobre um dos níveis. Neste caso, a probabilidade está difundida sobre todos os níveis, isto é, o nível k está afectado

pela probabilidade binomial $\binom{n}{k}\mu^k(1-\mu)^{n-k}$, onde $\mu=\frac{m}{n}$, $0 \le k \le n$. Por outro lado, esta probabilidade para o nível k é uniformemente repartida pois tudo se passa como no modelo anterior. Assim, primeiro escolhe-se o nível e depois um elemento desse nível.

N_3 : Este modelo aleatório consiste em:

Primeiro associar a E um conjunto aleatório E' do qual, somente se especifica a lei da variável aleatória $\upsilon = card(E')$, Poisson de parâmetro n:

$$P(\upsilon = l) = \frac{n^l}{l!}e^{-n}.$$

Seja $E' = E_0$ de cardinal l_0. O segundo passo consiste na escolha aleatória de um nível associado a D, no conjunto de partes de E_0. Esta escolha faz-se segundo o modelo binomial:

$$P(K=k) = \binom{l_0}{k}\mu^k(1-\mu)^{l_0-k}, \text{ onde } \mu=\frac{m}{l_0} \text{ e } k = 1,2,\ldots,l_0.$$

Seja $E' = E$ de cardinal l_0 e $K = k$ então, a escolha aleatória de X faz-se uniformemente sobre o nível k das partes de E.

Neste trabalho no entanto, só iremos utilizar o modelo N_1.

Seja então, π' uma partição aleatória de $P(n; t)$, obtida de acordo com uma distribuição de probabilidade uniforme sobre todas as partições de $P(n; t)$. Assim, ao índice $s(W,\pi)$ pode ser associada uma variável aleatória:

$$S(W,\pi') = card[gr(W) \cap (S(\pi') \times R(\pi'))]. \quad \textbf{(3.3.)}$$

A seguir, Lerman (1981), analisa a distribuição desta variável aleatória.

Vamos agora transcrever os principais resultados.

O índice (3.2.) pode ser escrito, na forma

$$s(W,\pi) = \sum\{card[gr(W) \cap (S(\pi) \times \{p\})] : p \in R(\pi)\}. \quad \textbf{(3.4.)}$$

Seja $k(p)$ a "posição" de p relativamente à ordem W e $h(p)$ a posição de p relativamente à restrição de W a $R(\pi)$:

$$k(p) = card\{q : q \in F \text{ e } q \leq p \text{ para } W\}$$
$$h(p) = card\{q : q \in R(\pi) \text{ e } q \leq p \text{ para } W\}.$$

Assim, para $p \in R(\pi)$, tem-se que:

$$card[gr(W) \cap (S(\pi) \times \{p\})] = k(p) - h(p)$$

e, portanto o índice (3.4.) assume a forma:

$$s(W,\pi) = \sum\{k(p) - h(p) : p \in R(\pi)\} = \sum\{k(p) : p \in R(\pi)\} - r(r+1)/2,$$

onde $r = card\{R(\pi)\} = \frac{1}{2}\sum_{j=1}^{k} n_j(n_j - 1)$.

Finalmente, o índice "bruto" pode ser escrito na forma:

$$s(W,\pi) = \sum\{\varepsilon(p)k(p) : p \in F\} - r(r+1)/2, \quad \textbf{(3.5.)}$$

onde $\{\varepsilon(p) : p \in F\}$ é a função indicatriz de R(π).

Analogamente, obtém-se para a variável aleatória (3.3.)

$$S(W,\pi') = \sum\{\varepsilon'(p)k(p) : p \in F\} - r(r+1)/2,$$

onde $\{\varepsilon'(p) : p \in F\}$ é a função indicatriz de R(π'). O que nos interessa é a primeira parte desta variável aleatória:

$$S(W,\pi') = \sum\{\varepsilon'(p)k(p) : p \in F\}.$$

Seja $f = card(F) = \binom{n}{2} = \frac{n(n-1)}{2}$.

Vamos agora apresentar dois resultados importantes de Lerman (1981):

Teorema 3.1.:

A média de $S(W,\pi')$ não depende do tipo de partição t e é $\frac{r(f+1)}{2}$, onde $r = card(R(\pi))$ e $f = card(F)$.

Teorema 3.2.:

Se as componentes de $t = (n_1, n_2, ..., n_k)$ forem iguais, a variância de $S(W,\pi')$ não depende do número de classes e é $rs(f+1)/12$, onde $s = f - r = card(S(\pi))$.

Lerman (1981, 1983) prova ainda, que na maior parte dos casos, a distribuição da variável aleatória $S_1(W,\pi')$ é assimptoticamente normal.

Estamos agora em condições de centrar e reduzir o índice "bruto" (3.5.):

$$\frac{[card(W) \cap (S(\pi) \times R(\pi))] - r \cdot s/2}{\sqrt{rs(f+1)/12}} = \frac{\sum\{\varepsilon(p)k(p) : p \in F\} - r \cdot s/2}{\sqrt{rs(f+1)/12}} \qquad \textbf{(3.6.)}$$

Este coeficiente pode ainda ser escrito de uma outra forma. Consideremos o conjunto de valores $\{k(p) : p \in F\} = \{1,2,...,f\}$.

A média e a variância deste conjunto de valores é $\frac{f+1}{2}$ e $\frac{(f-1)(f+1)}{12}$, respectivamente.

Seja

$$c(p) = \frac{k(p) - (f+1)/2}{\sqrt{(f^2-1)/12}}.$$

$c(p)$ mede a posição centrada e reduzida de p. Assim, o critério (3.6.) assume a forma

$$\frac{1}{\sqrt{r \times s/(f-1)}} \sum_{p \in F} \varepsilon(p) c(p). \quad \textbf{(3.7.)}$$

Obtemos assim, um critério de comparação entre uma partição do conjunto a classificar e o índice de semelhança inicial entre os elementos deste conjunto. Isto permite-nos escolher a melhor partição. Por exemplo, se tivermos uma partição em k classes e uma outra em k' classes, escolhemos aquela que maximiza o critério (3.7.).

O problema existente no critério (3.7.) é o volume de cálculo envolvido pois, o cálculo envolvido é de $O(n^2)$. Para pequenos conjuntos de dados isto pode não causar problemas; no entanto, nos dias de hoje, é cada vez mais frequente termos grandes conjuntos de dados que queremos dividir em classes. Por isso, Lerman reformulou o coeficiente (3.7.) de forma a que o volume de cálculo envolvido seja uma função linear de n (onde n é tamanho do conjunto de dados a classificar). Este novo coeficiente que pode ser utilizado com variáveis numéricas e categóricas, será descrito no capítulo seguinte.

3.2. Outros métodos

Seja E um conjunto de n objectos descritos por p variáveis e $d_{ii'}$ a distância entre os objectos i e i' (a escolha mais comum para $d_{ii'}$ é o quadrado da distância euclideana).

Suponhamos que temos os dados agrupados em k classes $C_1, C_2, ..., C_k$ e $n_r = |C_r|$.

Sejam $D_r = \sum_{i,i' \in C_r} d_{i,i'}$ a soma das distâncias de todos os pares de elementos da classe r e $W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r$.

Os procedimentos que procuram o valor mais apropriado para k são em geral referidos como sendo regras de paragem, porque perante a falta de informação acerca do valor ideal para k é necessário o uso de um algoritmo de classificação que permita obter uma partição óptima. Existem muitos métodos para estimar o número de classes; nesta secção vamos fazer um breve referência a alguns desses procedimentos.

Milligan e *Cooper* (1985) apresentaram os resultados de experiências de simulações desenvolvidas para testar 30 procedimentos diferentes. Nesta comparação foram utilizados 108 conjuntos de dados artificiais; para cada um dos conjuntos de dados foi calculada a matriz de dissemelhança (utilizando para tal a distância euclideana). Cada conjunto de dados foi analisado por quatro métodos de classificação. De todos os critérios usados o índice de *Calinski* e *Harabasz* (1974) foi o que produziu melhores resultados:

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

onde $B(k)$ e $W(k)$ são as somas dos quadrados das distâncias entre as classes e dentro das classes, respectivamente. A ideia consiste em escolher o k que maximiza $CH(k)$.

Krzanowski e *Lai* (1985) propuseram a quantidade $W_k k^{2/p}$ como um critério para a escolha do número óptimo de classes. Este método baseou-se noutro procedimento proposto por *Marriott* (1971). A actual proposta de *Krzanowski* e *Lai* (1985) define

$$DIFF(k) \doteq (k-1)^{2/p} W_{k-1} - k^{2/p} W_k$$

e escolhe o k que maximiza a quantidade

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

Esta maximização é semelhante à maximização de $W_k k^{2/p}$, mas os autores consideram que esta tem propriedades melhores. Note-se que, $KL(k)$ não está definida para k=1 e, por isso, não pode ser usada para testar apenas uma classe.

Hartigan (1975) propôs a estatística

$$H(k) = \left[\frac{W(k)}{W(k+1)} - 1 \right] (n-k-1),$$

a ideia era começar com k=1 e depois formar classes de forma a que $H(k)$ possa ser suficientemente grande. No entanto, *Hartigan* sugere que uma classe pode ser criada se $H(k) > 10$. Assim, o número óptimo de classes estimado é o mais pequeno $k \geq 1$, tal que, $H(k) \leq 10$. Esta estimação está definida para k=1.

Kaufman e *Rouseeuw* (1990) propuseram a estatística *Silhouette* para estimar o número óptimo de classes. Para observação i, seja $a(i)$ a média das distâncias desse ponto aos outros pontos dessa classe e $b(i)$ a média das distâncias aos pontos da classe mais próxima. Assim, a estatística *Silhouette* é definida como

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}.$$

Um ponto está bem classificado se $s(i)$ é grande. Estes autores propuseram a escolha do valor óptimo de número de classes $\hat{k}$ como o valor que maximiza a média de todo o conjunto de dados. Note-se que $s(i)$ não está definida para k=1.

Tibshirani, *Walther* e *Hastie* (2000) propuseram o método da estatística *Gap*. Definiram

$$Gap_n(k) = E_n^*(\log(W_k)) - \log(W_k),$$

onde E_n^* denota a esperança do conjunto de tamanho n da distribuição de referência. O valor óptimo para k é o que maximiza o critério $Gap_n(k)$. Assim a estimativa do número óptimo de classes é dado pelo valor de k para o qual $\log(W_k)$ está mais longe do seu valor esperado.

Muitos outros autores propuseram métodos de escolha do número de classes. Neste trabalho a partir de agora vamos só trabalhar com o critério de Lerman.

4. Extensões do algoritmo das *k*-médias

A propriedade mais atractiva do algoritmo das *k*-médias é, sem dúvida, a sua eficácia no processamento de grandes conjuntos de dados. No entanto, este algoritmo apenas trabalha com dados numéricos; este facto limita o seu uso em muitas aplicações onde intervêm dados categóricos. Foi com vista à resolução deste problema que, Huang (1997) apresentou dois novos algoritmos, que generalizam o método das *k*-médias, para dados com valores categóricos: o algoritmo das *k*-modas (Huang, 1997b) estende o método das *k*-médias para dados categóricos e o algoritmo dos *k*-protótipos (Huang, 1997a) integra os processos das *k*-médias e das *k*-modas para dados mistos (isto é, dados que apresentam simultaneamente variáveis numéricas e categóricas).

Seja $E = \{X_1, X_2, ..., X_n\}$ um conjunto com n objectos. Suponhamos que cada um dos elementos do conjunto E é descrito por p atributos, $A_1, A_2, ..., A_p$. Seja $Dom(A_i)$ o domínio de valores que o atributo A_i pode assumir.

Assim, iremos representar o objecto X_i pelo vector $(x_{i,1}, x_{i,2}, ..., x_{i,p})$. Todos os objectos têm exactamente p valores de atributos, isto é, não vamos permitir que falte nenhum atributo.

Dado um conjunto de objectos numéricos E e um número inteiro k $(\leq n)$ o algoritmo das *k*-médias procura uma partição de E, em k classes, que minimize um valor que resulta da soma das distâncias entre cada objecto e o centro do grupo a que estão actualmente alocados. Este processo é muitas vezes formulado como o problema P (Selim e Ismail, 1984; Bobrowski e Bezdek, 1991):

$$\text{Minimizar } P(W,Q) = \sum_{l=1}^{k}\sum_{i=1}^{n} w_{i,l}\, d(X_i, Q_l) \qquad \textbf{(4.1)}$$

tal que, $\sum_{l=1}^{k} w_{i,l} = 1$, $1 \leq i \leq n$ e $w_{i,l} \in \{0,1\}$, $1 \leq i \leq n$, $1 \leq l \leq k$.

Onde W é uma matriz partição $n \times k$, $Q = \{Q_1, Q_2, ..., Q_k\}$ é um conjunto de objectos do mesmo domínio e $d(.,.)$ é o quadrado da distância euclideana entre dois objectos.

O problema P pode ser resolvido iterativamente solucionando os dois problemas seguintes:

1. Problema P_1: Fixa-se $Q = \hat{Q}$ e a solução reduz-se ao problema $P(W, \hat{Q})$.
2. Problema P_2: Fixa-se $W = \hat{W}$ e a solução reduz-se ao problema $P(\hat{W}, Q)$.

O problema P_1 é resolvido por

$$w_{i,l} = 1 \text{ se } d(X_i, Q_l) \leq d(X_i, Q_t), \text{ para } 1 \leq t \leq k$$
$$w_{i,l} = 0 \text{ para } t \neq l$$

e o problema P_2 por

$$q_{l,j} = \frac{\sum_{i=1}^{n} w_{i,l} x_{i,j}}{\sum_{i=1}^{n} w_{i,l}}$$

para $1 \leq l \leq k$ e $1 \leq j \leq p$.

O algoritmo que resolve o problema P é dado por (Selim e Ismail, 1984; Bobrowski e Bezdek, 1991):

1. Escolhe-se um Q^0 inicial e resolve-se $P(W, Q^0)$ para obter W^0. Seja t=0.
2. Seja $\hat{W} = W^t$ e resolve-se $P(\hat{W}, Q)$ e obtém-se Q^{t+1}. Se $P(\hat{W}, Q^t) = P(\hat{W}, Q^{t+1})$, toma-se $\hat{W}$, Q^t e para-se; senão, ir para 3.
3. Seja $\hat{Q} = Q^{t+1}$ e resolve-se $P(W, \hat{Q})$ para obter W^{t+1}. Se $P(W^t, \hat{Q}) = P(W^{t+1}, \hat{Q})$, toma-se W^t, $\hat{Q}$ e para-se; senão, Seja $t = t$+1 e volta-se a 2.

Como $P(.,.)$ é não convexo e a sequência $P(.,.)$ gerada pelo algoritmo é estritamente decrescente, depois de um número finito de iterações o algoritmo converge para um mínimo local (Selim e Ismail, 1984).

4.1. O algoritmo das *k*-modas

O que impede o uso do algoritmo das *k*-médias para objectos categóricos é a medida de dissemelhança utilizada e o método usado para resolver o problema P_2. No entanto, estas barreiras podem ser ultrapassadas se fizermos as seguintes modificações no algoritmo das *k*-médias:

1. utilizar uma medida simples de dissemelhança para objectos categóricos (Kaufman e Rousseeuw, 1990);
2. usar as modas dos grupos em vez das médias;
3. usar um método baseado na frequência para encontrar as modas das classes.

4.1.1. Medida de dissemelhança

Seja X, Y dois objectos categóricos descritos por p atributos categóricos. A medida de dissemelhança entre X e Y pode ser definida através do número de diferenças entre os correspondentes atributos. Formalmente,

$$d_1(X,Y) = \sum \delta(x_j, y_j), \qquad \textbf{(4.2.)}$$

onde $\delta(x_j, y_j) = \begin{cases} 0 \text{ se } x_j = y_j \\ 1 \text{ se } x_j \neq y_j \end{cases}$.

4.1.2. Moda de um conjunto

Seja $X = \{X_1, X_2, ..., X_n\}$ um conjunto de objectos categóricos descritos por p atributos categóricos, $A_1, A_2, ..., A_p$.

Definição 4.1.:

A moda de X é um vector $Q' = (q_1, q_2, .., q_p)$ que minimiza $D(X, Q') = \sum_{i=1}^{n} d_1(X_i, Q')$.

Nota: Q' não é necessariamente um elemento de X.

Seja $n_{c_{k,j}}$ o número de objectos que têm como k-ésima categoria o atributo A_j e $fr(A_j = c_{k,j} \mid X) = \frac{n_{c_{k,j}}}{n}$ a frequência relativa da categoria $c_{k,j}$ em X.

Teorema 4.1.:

A função $D(X, Q)$ tem um mínimo se e só se $fr(A_j = q_j \mid X) \geq fr(A_j = c_{k,j} \mid X)$ para $q_j \neq c_{k,j}$ e para todo $j = 1, ..., p$.

Demonstração:

$$\sum_{i=1}^{n} d_1(X, Q') = \sum_{i=1}^{n} \sum_{j=1}^{p} \delta(x_{i,j}, q_j)$$

$$= \sum_{j=1}^{p} \left(\sum_{i=1}^{n} \delta(x_{i,j}, q_j) \right)$$

$$= \sum_{j=1}^{p} n\left(1 - \frac{n_{q_j}}{n}\right)$$

$$= \sum_{j=1}^{p} n\left(1 - fr\left(A_j = q_j \mid X\right)\right)$$

porque $n\left(1 - fr\left(A_j = q_j \mid X\right)\right) \geq 0$ para $1 \leq j \leq p, \sum_{i=1}^{n} d_1\left(X_i, Q'\right)$ tem um mínimo se e só se todo $n\left(1 - fr\left(A_j = q_j \mid X\right)\right)$ é mínimo. Então, $fr\left(A_j = q_j \mid X\right)$ é máximo.

O Teorema 4.1. define uma forma de encontrar Q' para um dado X; este resultado é importante porque permite que o método das *k*-médias seja usado em dados categóricos.

4.1.3. Algoritmo

Quando utiliza-se (4.2.) como medida de dissemelhança para os objectos categóricos, o custo da função (4.1.) é

$$P(W, Q') = \sum_{l=1}^{k} \sum_{i=1}^{n} \sum_{j=1}^{p} w_{i,l} \delta\left(x_{i,j}, q_{l,j}\right)$$

onde $w_{i,l} \in W$ e $Q_l = \left(q_{l,1}, q_{l,2}, \ldots, q_{l,p}\right) \in Q$.

1. Selecciona-se as *k*-modas iniciais, um para cada classe.

2. Na iteração número $l+1$, depois de calcular $q_1(l), q_2(l), \ldots, q_k(l)$, para cada $i = 1,2,\ldots,n$, reclassifica-se X_i na classe cuja moda está mais próxima e, logo de seguida, reactualiza-se o centro das duas classes.

3. Se não houver mudanças, para-se; senão voltar a 2.

Ainda não foi provado a convergência deste algoritmo (Anderberg, 1973). No entanto, na prática supõe-se que converge sempre.

O algoritmo das *k*-modas, tal como o das *k*-médias, apenas consegue encontrar um óptimo local. Este mínimo depende das modas iniciais tomadas e da ordem dos objectos do conjunto de dados.

4.2. O algoritmo dos *k*-protótipos

O algoritmo dos *k*-protótipos integra o algoritmo das *k*-médias e o das *k*-modas. Este algoritmo é na prática o mais útil, porque frequentemente os objectos são mistos.

4.2.1. Medida de dissemelhança

A medida de dissemelhança entre dois objectos X e Y, descritos pelos atributos $A_1^r, A_2^r, \ldots, A_m^r, A_{m+1}^c, \ldots, A_p^c$, pode ser definida por:

$$d_2(X,Y) = \sum_{j=1}^{m} (x_j - y_j)^2 + \gamma \sum_{j=p+1}^{p} \delta(x_j, y_j), \quad \textbf{(4.3.)}$$

onde o primeiro termo é o quadrado da distância euclideana (medida de dissemelhança para os atributos numéricos) e o segundo termo é a medida de dissemelhança para os atributos categóricos. O peso γ é usado para evitar favorecer um dos dois tipos de atributo. A influência de γ, no processo de classificação, está apresentada em (Huang, 1997a).

Utilizando (4.3.) para os objectos mistos, podemos modificar a função de custo (4.1) do seguinte modo:

$$P(W,Q)=\sum_{l=1}^{k}\left(\sum_{i=1}^{n}w_{i,l}\sum_{j=1}^{m}\left(x_{i,j}-q_{l,j}\right)^2+\gamma\sum_{i=1}^{n}w_{i,l}\sum_{j=p+1}^{p}\delta\left(x_{i,j},q_{l,j}\right)\right).$$

Seja $P_l^r=\sum_{i=1}^{n}w_{i,l}\sum_{j=1}^{m}\left(x_{i,j}-q_{l,j}\right)^2$ e $P_l^c=\gamma\sum_{i=1}^{n}w_{i,l}\sum_{j=m+1}^{p}\delta\left(x_{i,j},q_{l,j}\right)$.

Logo $P(W,Q)=\sum_{l=1}^{k}\left(P_l^r+P_l^c\right)$.

Como P_l^r e P_l^c são não negativas, minimizar $P(W,Q')$ é equivalente a minimizar P_l^r e P_l^c para $1\le l\le k$.

Dado um $\hat{Q}$ usa-se (4.3.) para calcular W como no algoritmo das *k*-médias. Dado um $\hat{W}$, encontra-se Q' minimizando P_l^r e P_l^c para $1\le l\le k$.

Um dos problemas no algoritmo dos *k*-protótipos, analogamente ao que acontece no algoritmo das *k*-médias, é como descobrir qual é o valor óptimo para *k*. Assim, o que vamos fazer é implementar, neste algoritmo (implementação cedida pelo autor), a adaptação linear do critério de Lerman.

Definição 4.2.:

A inércia total de um conjunto de n pontos descritos por p variáveis numéricas é dada por $\sum_{i=1}^{n}p(x_i)d^2(x_i,g)$, onde $g=(g_1,g_2,...,g_j,...,g_p)$ é o centro de gravidade:

$g_j=\frac{1}{n}\sum_{i=1}^{n}x_{ij}$, $j=1,...,p$.

Em geral, $p(x_i)=\frac{1}{n}$, $i=1,...,n$ e assim,

$$\text{Inércia total}=\frac{1}{n}\sum_{i=1}^{n}d^2(x_i,g).$$

Facilmente se mostra que

$$\text{Inércia total} = \sum_{i=1}^{n}\sum_{k=i}^{n} p(x_i)p(x_k)d^2(x_i,x_k) = \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{k=i}^{n} d^2(x_i,x_k).$$

Portanto, tem-se que

$$\frac{1}{n}\sum_{i=1}^{n} d^2(x_i,g) = \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{k=i}^{n} d^2(x_i,x_k). \quad \textbf{(4.4.)}$$

Seja

$$\alpha = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{k=i}^{n} d^2(x_i,x_k)$$

e

$$\lambda^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{k=i}^{n} d^4(x_i,x_k) - \alpha^2 .$$

Tendo por base da equação (4.4.), tem-se que:

$$\alpha = \frac{1}{n}\sum_{i=1}^{n} d^2(x_i,g)$$

e

$$\lambda^2 = \frac{1}{n}\sum_{i=1}^{n} d^4(x_i,g) - \alpha^2 .$$

Suponhamos agora que o conjunto dos objectos está repartido em k classes, $C_1, C_2, \ldots, C_k$.

Seja g o centro de gravidade global e g_l o centro de gravidade da classe C_l, $l = 1, \ldots, k$.

Para cada $X_i \in C_l$, associamos:

$$e(X_i, g_l) = \frac{d(X_i, g_l) - \alpha}{\lambda}.$$

$e(X_i, g_l)$ mede o afastamento centrado e reduzido do elemento X_i em relação ao centro da classe a que pertence.

Consideramos, de seguida, a soma:

$$\sum_{1 \le l \le k} \sum_{X_i \in C_l} e(X_i, g_l). \quad \textbf{(4.5.)}$$

Seja

$$\frac{1}{\sqrt{r \times s / f}} \sum_{1 \le l \le k} \sum_{X_i \in C_l} e(X_i, g_l)$$

a redução de (4.5.), onde $r = \sum_{1 \le l \le k} n_l^2$, onde $n_l = |I_l|, 1 \le l \le k$; $s = n^2 - r$ e $f = n^2$.

Seja $E = \{X_i : 1 \le i \le n\}$ o conjunto de dados a classificar. Suponhamos que os elementos de E estão descritos por p variáveis. Seja $Dom(m)$ o conjuntos dos valores assumidos pela variável m.

Assim, para determinar o critério linear pretendido temos de calcular:

1. O protótipo do conjunto de dados a classificar:

$$g = [g_1, g_2, \ldots, g_h, \ldots, g_m, f_{m+1}, \ldots, f_{m+j}, \ldots, f_p],$$

onde $g_h = \frac{1}{n} \sum_{1 \le i \le n} x_i^r$, $1 \le h \le m$ e f_{m+j} é a moda de $Dom(m+j)$.

2. O vector distância do protótipo anterior é:

$$[d_2(X_1,g),\ldots,d_2(X_i,g),\ldots,d_2(X_n,g)],$$

onde $d_2(X_i,g)=\sum_{1\le h\le m}(x_{ih}-g_h)^2+\sum_{1\le j\le p-m}\delta\left(x_{i(m+j)},f_{m+j}\right)$.

3. A média e a variância das componentes do vector distância:

$$\alpha=\frac{1}{n}\sum_{i=1}^{n}d_2(X_i,g)$$

e

$$\lambda^2=\frac{1}{n}\sum_{i\in I}[d_2(X_i,g)]^2-\alpha^2, \text{ respectivamente.}$$

4. Suponhamos que I_l é o conjunto de índices dos elementos da classe *l*. Então, $C_l=\{X_i : i\in I_l\}$.

Nestas condições determinamos para cada uma das classes o respectivo protótipo:

$$g^l=\left[g_1^l,g_2^l,\ldots,g_m^l,f_{m+1}^l,\ldots,f_p^l\right].$$

Este é obtido de forma análoga ao que foi descrito em 1, só que em relação à classe C_l.

De seguida, para cada elemento X_i pertencente a C_l, calcula-se:

$$e\left(X_i,g^l\right)=\frac{d_2\left(X_i,g^l\right)-\alpha}{\lambda}.$$

5. Finalmente calcula-se o critério:

$$\frac{1}{\sqrt{r \times s/n^2}} \sum_{1 \le l \le k} \sum_{i \in I_l} e\left(X_i, g^l\right), \quad \textbf{(4.6.)}$$

onde $r = \sum_{1 \le l \le k} n_l^2$, $s = n^2 - r$, com $n_l = |I_l|$, $1 \le l \le k$.

O critério obtido permite-nos escolher a melhor partição para um conjunto de dados que pretendemos classificar. Assim, se tivermos duas partições, uma em k classe e outra em t classes, escolhe-se aquela que o minimiza o critério (4.6.).

Vamos agora testar o critério de Lerman realizando experiências no algoritmo *k*-Protótipo (implementação disponibilizada pelo autor).

5. Análise do critéro de Lerman

Neste capítulo vamos testar o critério de Lerman utilizando o programa *k*-protótipos, cuja implementação foi fornecida pelo autor. Para tal, iremos gerar dados artificiais, com uma organização conhecida e verificar se a estrutura existente nos dados é identificada. Assim, utilizaremos quatro conjuntos com elementos com dois atributos para podermos visualizar os resultados.

Primeiro Exemplo

Trata-se de um conjunto muito simples, com apenas 9 elementos, representa um caso ideal sem problemas (está representado na Fig. 5.1).

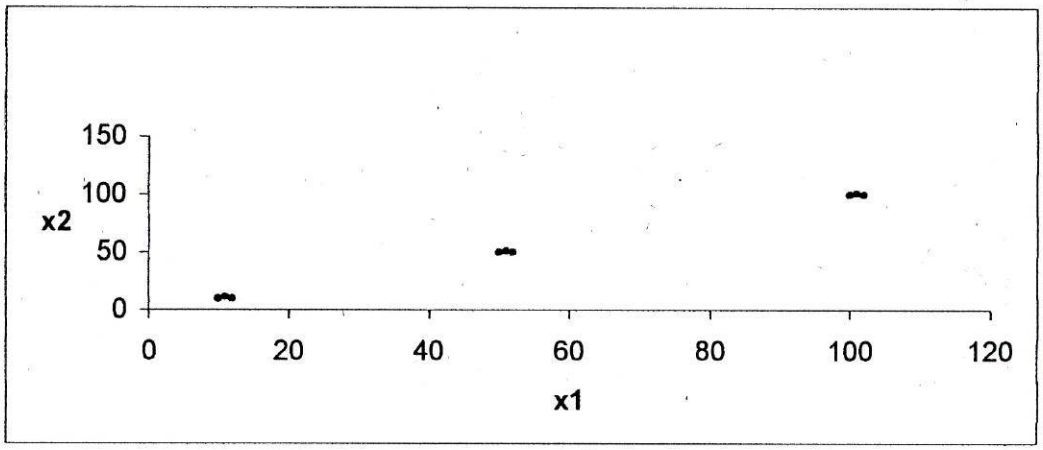

Fig. 5.1 Representação dos elementos do 1° exemplo

Observando o gráfico anterior, facilmente se conclui que o número óptimo de classes é 3.

Na tabela seguinte está representado o número de classes e o valor do critério de Lerman associado. Assim, na 1ª coluna está representado o número de classes, na 2ª coluna o valor do critério dando o mesmo peso a todas as classes e na 3ª coluna o valor do critério de Lerman.

Número de Classes	Valor do Critério de Lerman utilizando n/k	Valor do Critério de Lerman
2	-8,86477	-8,97550
3	**-8,99411**	**-8,99411**
4	-4,40866	-8,01014
5	-6,55036	-7,52536
6	-4,24416	-6,55625
7	-4,67883	-6,06462
8	-4,64245	-5,04352
9	-4,49853	-4,49853

Tab. 5.1 Valores do critério de Lerman para o 1° exemplo

Graficamente:

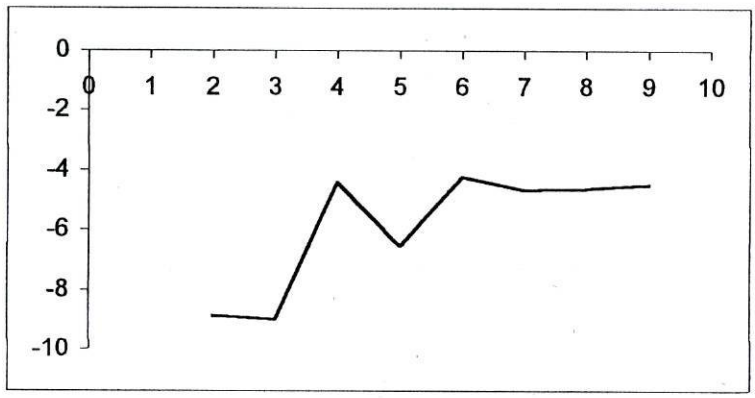

Fig. 5.2 Valor do Critério de Lerman utilizando n/k

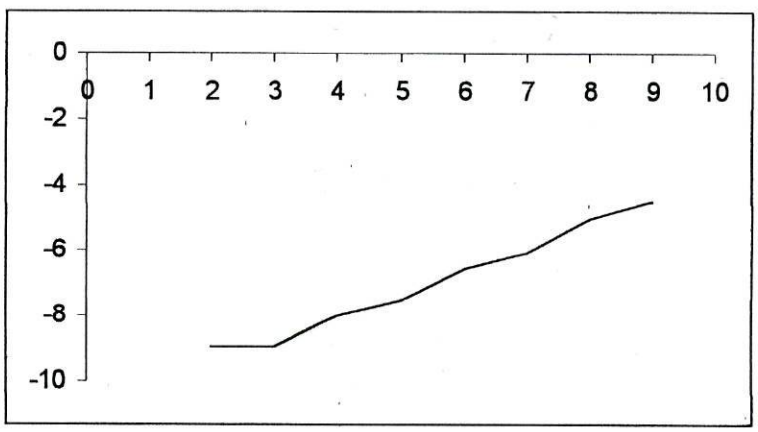

Fig. 5.3 Valor do Critério de Lerman

Observando a Fig. 5.1 facilmente se conclui que o valor óptimo do número de classes para este conjunto de dados é 3. Na Fig.5.2, podemos ver a evolução do critério de Lerman, enquanto o número de classes varia de 2 até 9. O valor mínimo deste critério Lerman é obtido para k=3.

Note-se, no entanto, que se em vez do critério de Lerman fosse utilizado o critério dando o mesmo peso a todas as classes, este também era mínimo para k=3, como se pode constatar observando a Fig. 5.3.

Segundo Exemplo

Um conjunto de 16 elementos que representa um caso sem problemas, como podemos constatar através do gráfico seguinte:

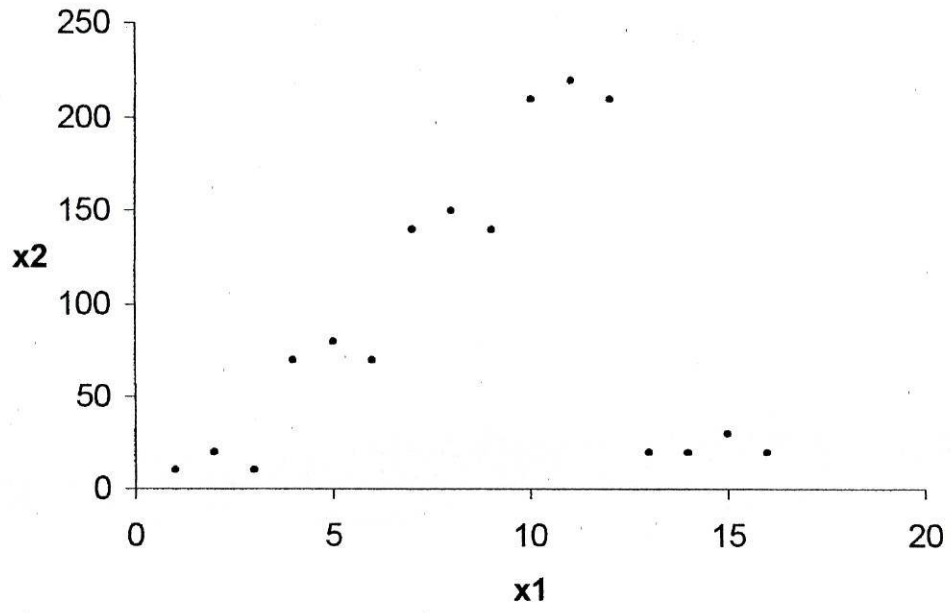

Fig. 5.4 Representação dos elementos do 2º exemplo

Neste caso, os resultados obtidos, para o critério de Lerman, estão na tabela seguinte:

Número de Classes	Valor do Critério de Lerman utilizando n/k	Valor do Critério de Lerman
2	**-15,15206**	**-16,30328**
3	-11,31689	-12,43137
4	-9,74581	-11,54897
5	-8,89408	-9,03332
6	-7,64543	-8,59866
7	-6,66236	-8,37631
8	-6,16917	-7,70790
9	-5,73172	-7,25047
10	-5,42635	-6,78022
11	-5,23099	-6,29326
12	-4,95452	-6,03800
13	-4,73969	-5,77616
14	-4,57725	-5,50667
15	-4,68334	-4,93944
16	-4,63755	-4,63755

Tab. 5.2 Valores do Critério de Lerman para o 2° exemplo

Graficamente:

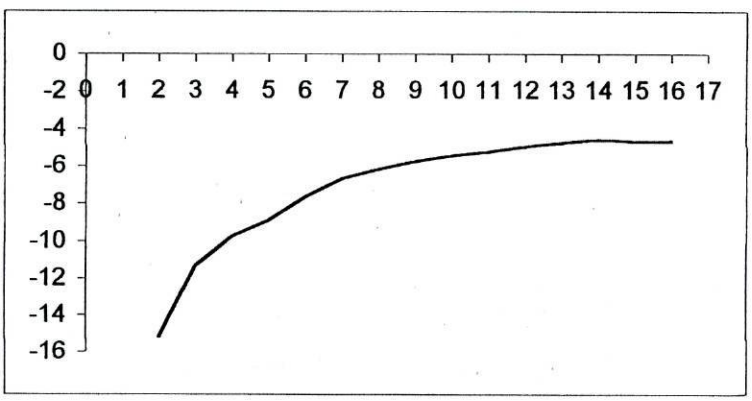

Fig. 5.5 Valores do Critério de Lerman utilizando n/k

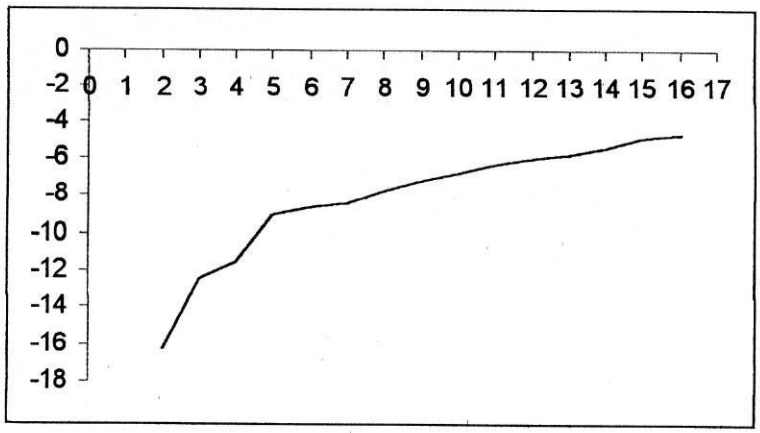

Fig. 5.6 Valores do Critério de Lerman

Neste caso, o critério de Lerman indica-nos k=2 como o valor óptimo para o número de classes. Se analisarmos a representação dos elementos (Fig. 5.4) verificamos que esta solução era de esperar para este conjunto de dados.

Terceiro Exemplo

Um conjunto com 21 elementos que reflecte a característica de possuir pontos isolados (está representado na Fig. 5.7).

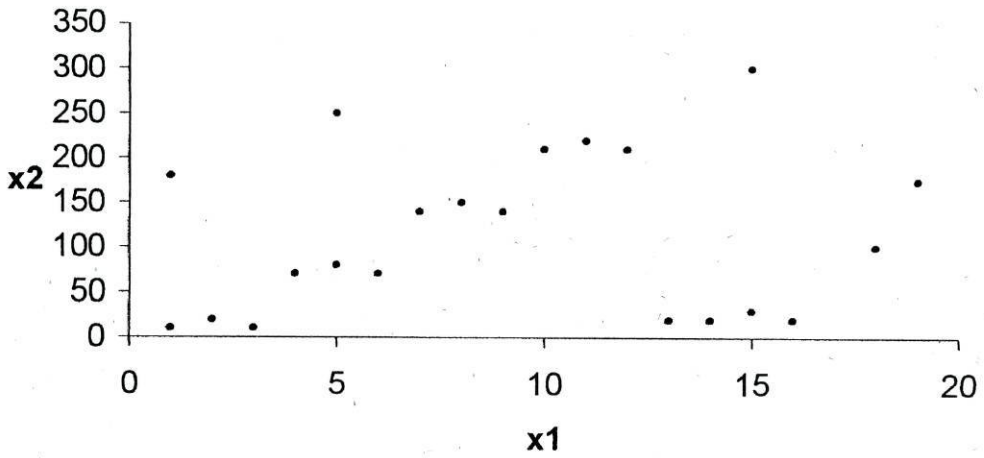

Fig. 5.7 Representação dos elementos do 3º exemplo

Neste caso, o valor do critério associado ao número de classes é:

Número de Classes	Valor do Critério de Lerman utilizando n/k	Valor do critério de Lerman
2	-17,12833	-21,98539
3	-13,48056	-17,54918
4	-10,76782	-16,22998
5	-8,82455	-14,37851
6	-8,64957	-12,20937
7	-7,62552	-11,23918
8	-6,92518	-10,85262
9	-6,22060	-10,26201
10	-6,47835	-9,85954
11	-6,07853	-9,34998
12	-5,66568	-8,78442
13	-5,52782	-8,34700
14	-5,35277	-8,11982
15	-5,23508	-7,65893
16	-5,17340	-7,17170
17	-5,01149	-6,91883
18	-4,88173	-6,65906
19	-4,78149	-6,39139
20	-4,89811	-6,12872
21	-4,87033	-4,87033

Tab. 5.3. Valores do Crietério de Lerman para o 3° exemplo

Graficamente:

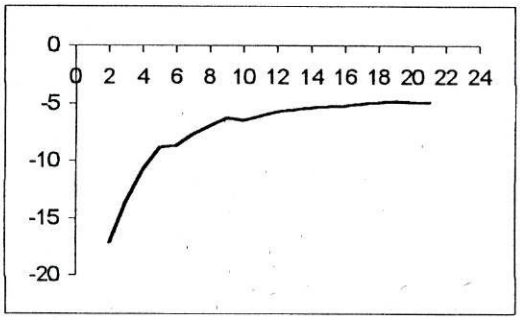

Fig. 5.8 Valos do Critério de Lerman utilizando n/k

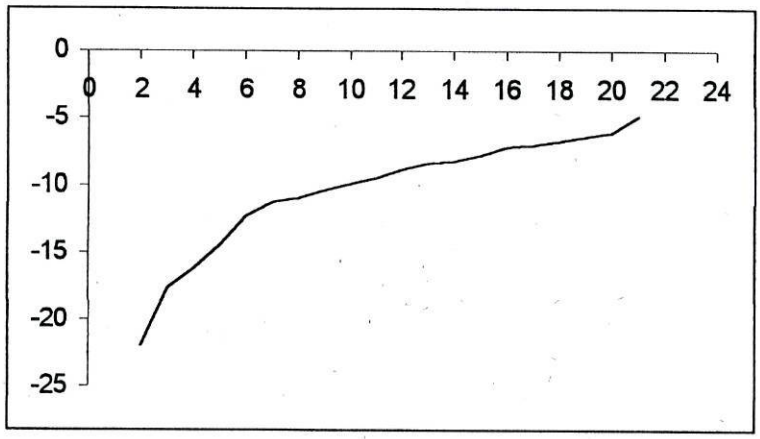

Fig. 5.9 Valores do Critério de Lerman

Neste exemplo é muito difícil prever qual é o número de classes ideal se considerarmos apenas a representação dos dados (figura 7), pois à partida várias soluções são possíveis. Na figura 8, podemos ver a evolução do critério de Lerman, quando o número de classes varia de 2 até 21. O valor mínimo deste critério Lerman é obtido para k=2.

Note-se, no entanto, que se em vez do critério de Lerman se utilizassemos o critério dando o mesmo peso a todas as classes, este também era mínimo para k=2, como se pode constatar observando a Fig.5.9.

Quarto Exemplo

Trata-se de um conjunto com 119 elementos que representa duas classes com grande disparidade de tamanho.

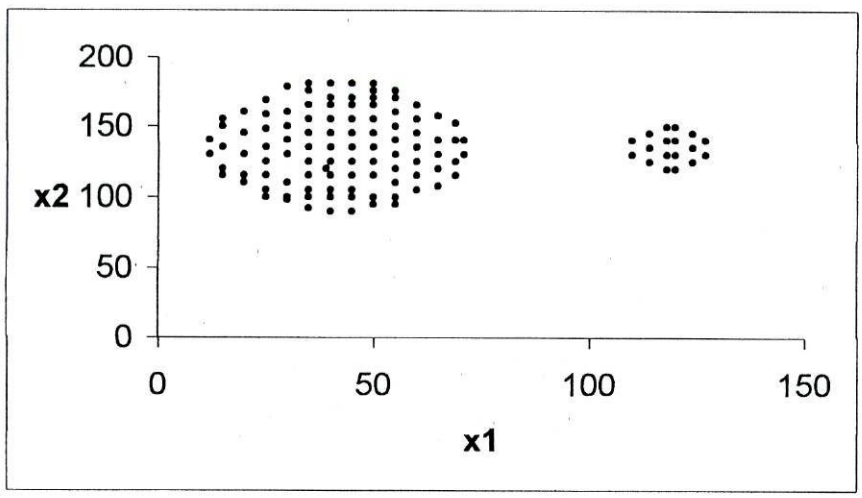

Fig. 5.10 Representação dos elementos do 4º exemplo

Neste caso, os valores do critério obtidos foram:

Nº de Classes	Valor do Critério de Lerman utilizando n/k	Valor do Critério de Lerman
2	**-43,15315**	-32,58232
3	-8,40493	-76,73704
4	-5,68052	-64,34994
5	-4,98236	-57,8767
6	-4,57305	-53,43407
7	-4,22316	-49,29978
8	-3,92261	-46,74844
9	-3,64352	-45,68704
10	-3,45541	-42,33568
11	-3,20893	-41,86584
12	-3,16363	-40,02161
13	-2,94032	-39,03002
14	-2,85434	-1005,25
15	-2,53262	-39,16425
16	-2,43324	-38,13227
17	-2,40913	-36,9855
18	-2,27529	-35,82541
19	-2,27936	-33,68071
20	-2,09289	-33,01785
21	-1,95184	-36,05281
22	-1,92264	-35,56118
23	-1,85385	-34,48988
24	-1,76711	-35,45304
25	-1,69944	-35,35931
26	-1,75188	-32,92401
27	-1,64783	-33,71391
28	-1,65124	-32,51041
29	-1,62652	-31,96293
30	-1,57409	-31,57085
31	-1,53921	-31,48263

32	-1,49813	-31,00143
33	-1,45102	-30,91926
34	-1,43044	-30,44834
35	-1,34311	-31,47400
36	-1,35097	-30,40748
37	-1,32914	-30,14908
38	-1,31237	-29,70219
39	-1,28537	-29,53027
40	-1,26654	-29,19129
41	-1,23730	-29,10655
42	-1,21935	-28,92533
43	-1,19923	-28,75506
44	-1,29623	-25,65400
45	-1,17929	-27,94382
46	-1,31028	-25,11938
47	-1,27249	-24,65358
48	-1,26062	-24,70324
49	-1,23005	-24,54067
50	-1,22727	-24,08438
51	-1,21312	-23,88388
52	-1,18979	-23,87418
53	-0,41692	-26,63012
54	-1,16769	-23,53161
55	-1,01273	-26,51133
56	-1,00127	-26,32880
57	-1,13605	-22,92299
58	-1,12159	-22,71117
59	-1,11288	-21,88038
60	-1,10477	-21,49386
61	-1,09723	-21,08889
62	-1,08219	-21,02937
63	-1,05981	-21,03760
64	-1,07240	-21,27648

66	-0,89256	-25,04938
67	-1,04623	-21,22351
68	-1,03364	-21,05421
69	-1,01047	-20,10894
70	-1,00139	-20,22370
71	-0,99538	-20,16188
72	-0,98426	-19,73554
73	-0,97618	-82,46754
74	-0,32800	-19,78070
75	-1,01153	-77,24646
76	-0,94561	-75,56733
77	-0,32778	-70,69613
78	-0,33466	-67,57616
79	-0,33865	-66,05144
80	-0,33877	-63,06812
81	-0,34382	-61,60788
82	-0,34451	-60,16758
83	-0,34543	-60,16758
84	-0,94643	-57,34398
85	-0,34791	-55,95938
66	-0,89256	-25,04938
86	-0,34948	-50,58887
87	-0,37055	-50,58887
88	-0,36633	-49,28565
89	-0,36912	-47,99728
90	-0,37220	-45,46357
91	-0,38341	-18,78721
92	-0,85517	-39,36540
93	-0,42014	-39,36540
94	-0,41567	-39,36540
95	-0,41130	-34,71589
96	-0,45154	-34,71589
97	-0,44689	-18,06826

99	-0,46351	-30,26676
100	-0,48754	-17,86036
101	-0,81948	-21,76171
102	-0,65942	-17,72077
103	-0,80978	-16,33974
104	-0,87280	-15,25611
105	-0,92178	-15,01924
106	-0,91308	-14,53355
107	-0,96802	-14,04997
108	-0,97700	-13,72285
109	-0,98687	-13,54761
110	-0,99098	-13,37018
111	-0,99555	-13,01012
112	-1,01518	-12,64025
113	-1,02130	-12,54267
114	-1,02009	-12,44435
115	-1,01907	-12,34528
116	-1,01827	-12,24544
117	-1,01767	-12,14679
118	-1,01727	-12,04737
119	-1,01709	-1,01709

Tab. 5.4 Valores do Critério de Lerman para o 4º exemplo

Graficamente:

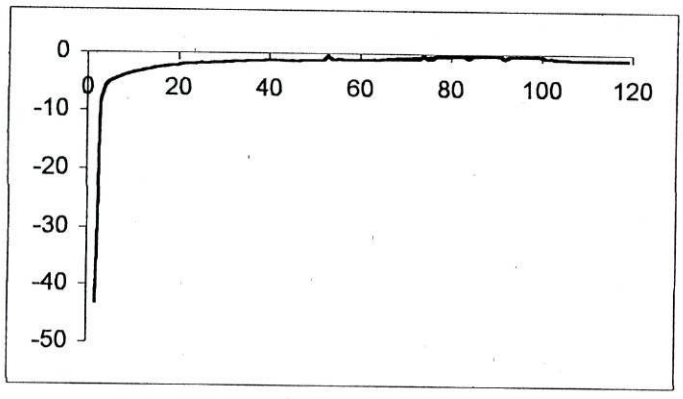

Fig. 5.11 Valores do Critério de Lerman utilizando n/k

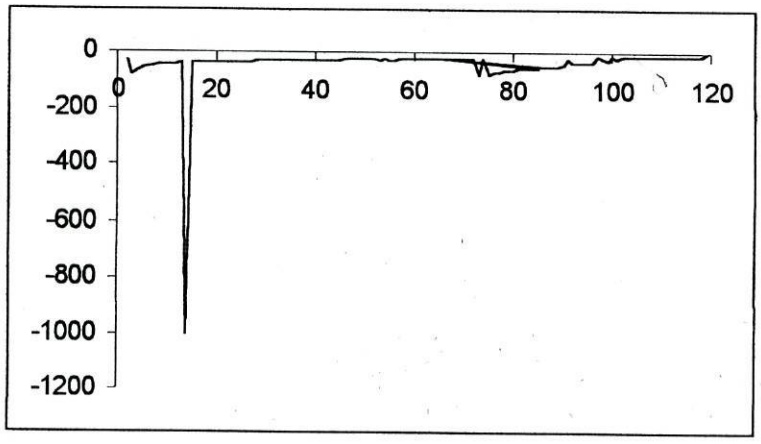

Fig. 5.12 Valores do Critério de Lerman

Observando a Fig. 5.11, facilmente se conclui que o número óptimo de classes é 2. Na Fig. 5.12, podemos observar a evolução do critério de Lerman dando o mesmo peso a todas as classes, para k=2, ..., 119. O valor mínimo é obtido para k=2.

No entanto, neste caso o critério de Lerman não é mínimo para k=2, mas sim para k=14. Isto deve-se ao facto do algoritmo dos k-protótipos distribuir, para k=14, os dados do conjunto apenas em duas classses: uma classe com 1 elemento e outra com 118 elementos; ou seja, as restantes 12 classes não têm qualquer elemento. Como os elementos não estão distribuídos de forma uniforme pelas classes e como no calcular o valor do critério de Lerman é necessário o cardinal das classes, o critério de Lerman assume por isso, um valor muito anormal para k=14.

Assim, para que possamos chegar ao valor óptimo do número de classes é preciso analisar os valores obtidos através do critério de Lerman, juntamente com os resultados obtidos pelo algoritmo dos k-protótipos.

6. Conclusão

Neste trabalho estudámos um dos métodos de agrupamento que permite, de um conjunto de observações, obter uma partição em k grupos – o método de classificação das k-médias de MacQueen(1967). Foi feita uma descrição deste método, das suas propriedades, bem como os diferentes problemas levantados por este método.

Como o objectivo principal era tratar a problemática da escolha do valor óptimo para k (ou seja, quantos grupos devem ser formados) foi dado um especial ênfase ao longo deste trabalho a esta temática.

Nesta dissertação foram também descritos alguns métodos já desenvolvidos, para determinar o número de classes existentes num conjunto de dados, e foi apresentado um novo método desenvolvido por Lerman. Por isso, grande parte deste trabalho foi destinado a descrever e a analisar este novo método, utilizando o algoritmo dos k-protótipos (cuja implementação foi disponibilizada pelo autor). O critério de Lerman e o algoritmo dos k-protótipos são completamente independentes, pelo que a informação que advém do critério de Lerman não tem qualquer influência no algoritmo dos k-protótipos.

Os resultados obtidos na análise realizada leva-nos a crer que o método de Lerman produz bons resultados; tanto no caso em que é dado o mesmo peso a todas as classes como no caso em que este critério utiliza o cardinal das classes. No entanto seria necessário uma análise mais profunda e detalhada para comparar ambos os casos.

Assim, o critério de Lerman é um novo método disponível, para escolher o número óptimo de classes, em Análise Classificatória.

7. Bibliografia e Referências

1. M. Abramowtiz and Stegun; *"Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables (Nat. Bur. Of Stand. Appl. Math. Ser., n°55)"*, 7° Priting. USGord. Priting Office, Washington, D.C., (1968).

2. M.R. Anderberg; *"Cluster analysis for applications"*, Academic Press, New York and London (1973).

3. G.H. Ball and D.J.Hall; *"ISODATA, a Novel Method of Data Analysis nad Pattern Classification"*, AD 699616, Stanford Res. Inst., Menlo Park, California (1965).

4. J.P. Benzecri; *"Analyse factorielle des proximités"*, I & II, Pubi. De l'Inst. De Stat. De l'Univ. de Paris, XIII & XIV, (1964-65).

5. J.C. Bezdek; *"Pattern Recognition with Fuzzy Objective Function"*, Plenum Press, (1981).

6. L. Bobrowski and J.C. Bezdek; *"c-Means clustering with the* l_1 and l_∞ *norms"*, IEEE Translations on Systems, Man and Cybernetics, 21(3): 545-554, (1991).

7. R. B. Calinski and J. Harabasz; *"A dendrite method for cluster analysis"*, Communications in statistics 3, 1-27, (1974).

8. J. F. Costa, I.C. Lerman and H. Silva; *"Linéarisation d'un Critère de Classification en Cas de Données Numériques et Qualitatives Nominales"*, "Article envoyé à la Conferènce: 8 èmes Rencontres de la Société Francophone de Classification", 17-21/12/01; Guadeloupe.

9. H.E. Daniels; *"The relation between measures of correlation in the universe of samples permutations"*, Biometrika, vol. 33, (1994).

10. E. W. Forgy; *"Cluster Analysis of Multivariate Data: Efficiency Versus Interpretability of Classifications"*, Biometric Soc. Meetings, Riverside, California, Abstract in Biometrics 21, no.3, 768, (1965).

11. C. Fraley and A. E. Raftery; *"How Many Clusters? Which Clustering Methods? AnsWers Via Model-Based Cluster Analysis"*, Technical Report n° 329, Department of Statistics, University of Washington, (1998).

12. K. Fukunaga; *"Introduction to Statistical Pattern Recognition"*, Second Edition, Academic Press, (1990).

13. A. Gordon; *"Classification (2nd edition)"*, Chapman and Hall /CRC press, London, (1999).

14. Z. Huang; *"A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining"*, Cooperative Research centre for Advanced Computacional Systems, Canberra, Austrália, (1998).

15. Z. Huang; *"Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values"*, Data Mining, Knowledge Discovery, Vol. 2, No. 3, 283-304, (1998).

16. Hartigan, J.; *"Classification (2^{nd} edition)"*, Chapman and Hall/CRC Press, London, (1975).

17. L.J. Hubert; *"Generalized proximity function comparations"*, Br. J. math. Static. Psychol., 31, 179-192, (1978).

18. L.J. Hubert, F B. Baker; *"Evaluating the conformity of sociometric measurements"*, Psychometrika, 43, 1, 31-41, (1978).

19. L. Kaufman and P. Rousseeuw *"Finding groups in data: an introduction to cluster analysis"*, New York, Wiley, (1990).

20. M.G. Kendall; *"Rank Correlation Methods"*, Charles Griffin, fourth edition, (1971).

21. W. J. Krzanowski and Y. T: Lai, *"A criterion for determining the number of groups in a data set using sum of squares clustering"*, Biometrics 44, 23-34, (1985).

22. G. Lecalve; *"Problèmes d'analyse des donnes"*, $2^{ème}$ partie d'une thèse d'état, Univ. Rennes I, Nov., (1976).

23. I.C. Lerman; *"Les bases de la classification automatique"*, Gauthier-Villars, Paris, (1970).

24. I.C. Lerman; *"Sur l'analyse des données préalable à une classification automatique; proposition d'une nouvelle mesure de similarité entre classes"*, Rev. Math. & Sc. Hum., n°32, (1970a).

25. I.C. Lerman; *"Etude distributionnelle de statistiques de proximité entre structures finies de même type; application à la classification automatique"*. Cahiers du B.U.R.O. n°19, Paris, (1973).

26. I.C. Lerman; *"Formal analysis of a general notion of proximity between variables"* in Proceed. (published by North Holland in 1977), Congrès Européen des Statisticiens, Grenoble, Sept., (1976).

27. I.C. Lerman; *"Classification et analyse ordinale des données"*, Dunod, 760 p., Paris, (1981).

28. I.C. Lerman; *"Corrélation partielle dans le cas qualitatif"*, rap. IRISA – Rennes n° 153, (à paraître dans les "Publ. De l'Inst. De Stat. De l'Univ. de Paris), (1981a).

29. I.C. Lerman, R. Gras, H. Rostam; *"Elaboration et évaluation d'un indice d'implication pour des donées binaires"* I et II, Rev. Math. Sc. Hum., 19 ème année n°74 et n° 75, Paris, (1967).

30. I.C. Lerman; *"Sur la signification des classes issues d'une classification automatique de donnes"*, NATO ASI Series, Vol. G1 Numerical Taxonomy, (1983).

31. J. MacQueen; *"Some Methods for Classification and Analysis of Multivariate Observations-in Proc. Of the 5th Berkeley Symposium on Mathematical Statistics and Probability"*, Vol.1, 281-297, (1967).

32. N. Mantel; *"The detection of disease clustering and a generalized regression approach"*; Cancer research, 27, 209-20, (1967).

33. F. Nicolau; *"Critérios de análise classificatória hierárquica baseados na função de distribuição"*, Laboratório de Estatística, Faculdade de Ciências de Lisboa, (1980).

34. F. Nicolau, P. Brito; *"Improvements in NHMEAN Method"*, in "Data Analysis, Learning Symbolic and Numerical Knowledge", Ed. E. Diday, New Science Publishers, Inc., New York, (1989).

35. G. W. Milligan and M. Cooper *"An examination of produces for Determining the Number of Clusters in a Data Set"*, Psychometrika, Vol. 50, n°. 2, pp. 159-179, Junho, (1985).

36. H. Ralambondriny; *"A conceptual version of the k-means algorithm"*, Pattern Recognition Letters, 16: 1147-1157, (1995).

37. E.R. Ruspini; *"A new approach to clustering"*, Information Control, 19: 22-32, (1969).

38. E.R. Ruspini; *"New experimental results in fuzzy clustering"*, Information Sciences, 6: 273-284, (1973).

39. R.N. Shepard; *"The analysis of proximities: multidimensional scaling with an unkown distance function"*; Psychometrika, vol.27, (1962).

40. P.H.A. Sneath; *"Some empirial testes based for significance of clusters"* in "Data Analysis and Informatics" E. Diday et al. (eds), North Holland, (1980).

41. R. Tibshirani, G. Walther and T. Hastie, *"Estimating the number of clusters in a dataset via the Gap statistic"*, 29 March, (2000).

42. D. Wishard; *"Fortran II Programs for 8 methods of cluster analysis (Clustan I)"*,Comput. Contrib. 38 State Geol. Survey. Univ. of Kansas, Lawrence.

43. Wald and J. Wolfowitz; *"Statistical tests based on permutations of the observations"*; Ann. Math. Stat. Vol. 15, 358-372.