CLASSIFICAÇÃO DE OBJECTOS COM CARACTERÍSTICAS PARCIALMENTE DESCONHECIDAS

Manuel Lameiras Campagnolo (*)
Helder Coelho (**)

1 — Introdução

Considere-se um conjunto de objectos caracterizados por um conjunto de variáveis. Uma característica é uma instanciação de uma variável que descreve o objecto. Em classificação não supervisada, ou *cluster analysis,* o objectivo é agrupar esses objectos em classes, por forma a que os objectos de uma classe sejam mais semelhantes entre si do que os objectos de classes distintas.

Existem vários tipos de métodos para agrupar os objectos de características conhecidas. Segundo Gnanadesikan *et al.* (1989), podem ser distinguidos três tipos de métodos numéricos para a classificação não supervisada: métodos hierárquicos; métodos de divisão e métodos de sobreposição. Os métodos numéricos envolvem a utilização de uma distância (ou medida de semelhança) entre os objectos.

Existem igualmente métodos não automáticos que consistem na redução da dimensionalidade do problema e na análise gráfica dos dados em duas ou três dimensões.

Os métodos hierárquicos serão aqui tomados como principal referência. Estes são os mais utilizados no domínio das aplicações devido às suas características de simplicidade e disponibilidade (v. Gnanadesikan *et al,* 1989). Nesses métodos é definida uma distância entre conjuntos de objectos e são construídas hierarquicamente classes por forma a minimizar a distância entre objectos da mesma classe e maximizar a distância entre objectos de classes distintas.

Suponhamos agora que, para cada objecto, é conhecido apenas um subconjunto de variáveis. Nesse caso a definição de uma distância entre objectos não é uma questão trivial porque nem todas as características dos objectos são conhecidas. O problema que se põe, portanto, é o de construir uma medida de semelhança para objectos de características parcialmente desconhecidas. Definida essa medida a classificação pode então ser realizada da forma convencional.

Para resolver o problema propõe-se uma abordagem em duas fases. Na primeira fase complementa-se a informação existente sobre cada objecto estimando distribuições de probabilidade para as variáveis associadas às características desconhecidas do objecto. Esta estimação é realizada a partir do conhecimento de relações entre as variáveis envolvidas e é suportada por um formalismo de incerteza. Na segunda fase é construída uma medida de semelhança entre os objectos completa e incertamente caracterizados na fase anterior.

(*) Departamento de Matemática, ISA/UTL.
(**) Núcleo de Tecnologias e Ciência da Informação, ISEG/UTL.

Em 2 será definido formalmente o problema, em 3 será descrita a forma de estimar as características desconhecidas dos objectos e em 4 a forma de construção da medida de semelhança. Em 5 será descrito sumariamente o algoritmo para agrupamento dos objectos em classes. Em 6 será dado um exemplo de aplicação da metodologia proposta.

2 — Definição formal do problema e notações

Considere-se que um objecto é descrito por m variáveis, $X_1, \ldots, X_m$, quantitativas ou qualitativas. Considere-se também que cada variável admite um número finito de modalidades. O número de modalidades da variável X_k será denotado por n_k. Seja x_k a modalidade que toma a variável X_k para um determinado objecto. Esse objecto pode portanto ser representado por um vector $x = (x_1, \ldots, x_k, \ldots, x_m)$ que pode ser designado vector de características do objecto. A k-ésima características do objecto será, portanto, a modalidade a que pertence o valor da k-ésima variável que descreve esse objecto.

As notações atrás definidas são insuficientes para a representação de objectos cujas m características não são completamente conhecidas. Para tratar esse caso mais geral será considerada uma formalização análoga mas de natureza probabilística. Em alternativa a tomar-se a k-ésima característica do objecto como conhecida, considerar-se-á que existe uma distribuição de probabilidade das modalidades da k-ésima variável que será representada por p_k ou por $p(x_k)$. Seja então $\{x_{k1}, \ldots, x_{kn_k}\}$ o conjunto de modalidades da variável X_k. A probabilidade de, para um determinado objecto, a variável X_k tomar a modalidade x_{kj} será p_{kj}. Um objecto pode, neste formalismo, ser representado por:

$$x = ((p_{11}, p_{12}, \ldots, p_{1n1}), \ldots, (p_{k1}, p_{k2}, \ldots, p_{knk}), \ldots, (p_{m1}, p_{m2}, \ldots, p_{mnm})) \text{ com } \Sigma_{j=1}^{nk} p_{kj} = 1, \forall_k.$$

Um objecto que seja descrito por quatro variáveis, respectivamente com quatro, três, quatro e duas modalidades e cujas características sejam completamente conhecidas e será representado, por exemplo, por $x = ((0, 0, 1, 0), (0, 1, 0), (1, 0, 0, 0), (1, 0))$, e um objecto semelhante cuja segunda característica seja desconhecida será representado por $x = ((0, 0, 1, 0), \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right), (1, 0, 0, 0), (1, 0))$, significando que a probabilidade de a segunda variável tomar qualquer uma das suas modalidades é idêntica. Dizer-se-á que um objecto é conhecido relativamente à k-ésima variável se a distribuição p_k para esse objecto for da forma $\exists_i: p_{ki} = 1$ e $p_{kj} = 0$, $\forall_{j \neq i}$.

Sempre que seja necessário as notações atrás definidas serão completadas com a identificação do objecto. Por exemplo, $p_{kj}(x)$ representará a probabilidade p_{kj} para o objecto x.

Por evidência designa-se uma determinada instanciação de um subconjunto de variáveis de $\{X_1, \ldots, X_m\}$. Relativamente ao problema aqui tratado, a noção de evidência é uma noção central. Considere-se um objecto de características parcialmente desconhecidas. A evidência existente sobre esse

objecto é o conjunto das características conhecidas. Essa evidência pode ser utilizada para inferir algum conhecimento sobre as características desconhecidas. Em 3 este último aspecto será desenvolvido.

A resolução do problema consiste em definir classes que contenham os objectos a classificar por forma a que um objecto de uma classe seja mais semelhante aos restantes objectos da sua classe do que aos objectos de qualquer outra classe. Considera-se que as classes são disjuntas. Cada classe é, portanto, um conjunto de objectos $C = \{x_1, \ldots, x_q\}$ tal que, para $x_i \in C$, a semelhança entre x_i e $C - \{x_i\}$ é maior do que entre x_i e a classe C', qualquer que seja $C' \neq C$. É necessário, portanto, definir uma medida de semelhança entre um objecto x e um conjunto de objectos C. Essa medida será representada por $S(x, C)$ e será definida a partir de uma medida de semelhança entre dois objectos (x_1 e x_2), que será representada por $s(x_1, x_2)$. Assim, $s(x_1, x_2)$ será definida em 4, e $S(x, C)$ será definida de uma das formas convencionais para métodos hierárquicos de classificação que serão apresentadas em 5.

O problema considerado neste trabalho pode ser enunciado formalmente do modo seguinte. Dado um conjunto de objectos $\{x_1, \ldots, x_n\}$ de características parcialmente desconhecidas, definir N classes, $C_1, \ldots, C_N$, que constituam uma partição do conjunto de objectos, tais que:

$$\forall_{x \in ci},\ \forall_{j \neq i},\ S(x,\ C_i - \{x\}) \geq S(x,\ C_j)$$

3 — Complementação da caracterização dos objectos

Dado que para um determinado objecto se admite que existem características desconhecidas, tem interesse poder obter algum conhecimento sobre essas características. Neste trabalho sugere-se que esse conhecimento possa ser representado por uma estimação de uma distribuição de probabilidades para cada variável desconhecida.

Essa estimação pode ser realizada com base no conhecimento das características conhecidas do objecto e através do conhecimento das relações existentes entre a totalidade das variáveis que descrevem o objecto. Supõe-se, portanto, a existência desse tipo de relações.

Por um lado, é necessário utilizar algum formalismo que suporte essas relações e que permita explorá-las. O conceito de grafo bayesiano responde a esta necessidade. Por outro lado, deve ser analisada a forma de estabelecer e quantificar essas relações. Para este aspecto serão considerados o conhecimento que se pode retirar dos dados disponíveis e o eventual conhecimento pericial existente sobre o domínio dos objectos.

O grafo bayesiano, também designado por rede de crenças [ou por *belief network, influence network* (Pearl, 1987) ou *probabilistic influence diagram* (Geiger *et al.*, 1990)], é formalmente, um grafo orientado acíclico construído a partir de uma distribuição de probabilidade de um conjunto de variáveis aleatórias, $G = (G, p)$, em que G representa o grafo orientado acíclico e p a distribuição de probabilidade.

O grafo bayesiano é constituído pelas duas seguintes componentes. A primeira designa-se componente gráfica e consiste num modelo gráfico de relações de dependência entre as variáveis. Este modelo representa o conhecimento qualitativo sobre o domínio do problema pois apenas exprime a existência de relações de dependência entre as variáveis envolvidas, sem asso-

ciar uma medida de grandeza a essas relações. A segunda componente designa-se componente probabilística e caracteriza quantitativamente as relações expressas pelo modelo gráfico. Designa-se probabilística porque se baseia na teoria das probabilidades para associar medidas de incerteza ao modelo de relações entre as variáveis. A componente probabilística do grafo bayesiano permite, conjuntamente com a componente gráfica, estimar a distribuição de probabilidade, para cada uma das variáveis não instanciadas, dada uma evidência. Um exemplo concreto de grafo bayesiano é apresentado em 6. A componente gráfica e a componente probabilística desse grafo bayesiano são descritas na figura 1 e na tabela 4, respectivamente.

O grafo bayesiano é uma estrutura que tem sido alvo de grande interesse pois permite representar de uma forma muito atraente a organização do conhecimento humano que se baseia no estabelecimento de um conjunto de proposições (representadas pelos vértices do grafo) que descrevem a realidade e na pesquisa e actualização do conhecimento por sequências de inferências sobre as relações relevantes entre essas proposições (relações que são representadas pelas arestas do grafo). O grafo bayesiano pode ser visto não apenas como uma estrutura de representação do conhecimento, mas também como uma arquitectura computacional para raciocinar sobre esse conhecimento (Pearl, 1986).

Neste texto, G denotará o grafo orientado constituído pelo conjunto de vértices $X = \{X_1, \ldots, X_m\}$ que correspondem ao conjunto de variáveis que descrevem os objectos a classificar e pelo conjunto de arestas orientadas que representam, cada uma, a existência de uma relação directa entre um par de variáveis de X. Seja (X_i, X_j) a aresta orientada que representa a relação de dependência da variável X_j relativamente à variável X_i. O conjunto de variáveis X_i para as quais existe uma aresta (X_i, X_k) será designado o conjunto dos pais de X_k e será representado por $\text{pa}(X_k)$. O termo «família de X_k» identificará o conjunto de variáveis $\{X_k\} \cup \text{pa}(X_k)$.

A distribuição $p(x_k)$, definida em 2, também pode ser designada como distribuição de probabilidade marginal de X_k se se considerar que existe uma distribuição conjunta para $X_1, \ldots, X_m$. Seja $p(x) = p(x_1, \ldots, x_m)$ a distribuição de probabilidade conjunta e $p(x_k | A)$, com $A \subset X$, a distribuição da probabilidade de X_k condicionada por uma instanciação das variáveis de A, ou seja, por uma evidência. A propagação de uma evidência pelo grafo é realizada por uma série de processadores locais que correspondem, cada qual, a uma variável e aos pais dessa variável. Para propagar uma evidência pelo grafo é necessário conhecer, para cada variável, a distribuição de probabilidade dessa variável condicionada pelos seus pais, isto é, para X_k, $p(x_k | pa(X_k))$.

Existem diversos métodos de propagar uma evidência pelo grafo bayesiano que se baseiam todos no seguinte resultado: «A distribuição conjunta das variáveis $X_1, \ldots, X_m$ é igual ao produto das distribuições de probabilidade de cada uma condicionadas pelos seus pais, isto é, $p(x) = \prod_{i=1}^{m} p(x_i | pa(X_i))$.» Este aspecto não será aqui desenvolvido mas poderá ser aprofundado em Pearl (1987), Spiegelhalter *et al.* (1992) e Pearl (1986). Uma síntese e uma comparação dos diversos métodos poderá ser encontrada em (Campagnolo, 1992).

Qualquer que seja o método de propagação o objectivo é determinar uma estimativa da distribuição de probabilidade das variáveis cujo valor é desconhecido a partir de uma evidência. No problema de classificação considerado neste trabalho isto permite, para cada objecto, determinar uma distribuição de probabilidade para todas as variáveis desconhecidas a partir das características conhecidas e assim completar a caracterização do objecto.

Um problema que evidentemente se coloca com acuidade é o de construir o modelo gráfico das relações de dependência entre as variáveis. Esse problema pode ser resolvido com base na informação dos dados (v. Whittaker, 1990) ou, de uma forma simples recorrendo ao conhecimento pericial existente sobre o domínio do problema. Esse conhecimento consistiria apenas no estabelecimento de relações de dependência entre as variáveis envolvidas.

Outro problema que se põe é o de determinar as distribuições de probabilidade condicionadas necessárias. Este problema pode ser resolvido recorrendo a dois tipos de informação. Por um lado, está disponível a informação dos dados, ou seja, as características conhecidas dos objectos. Dado que as probabilidades condicionadas a estimar respeitam apenas o conjunto de variáveis $\{X_k\} \cup pa(X_k)$, para cada k, uma forma de proceder à estimação é por determinação das frequências, para os objectos em que todas essas variáveis estão instanciadas, da ocorrência simultânea das características associadas a $\{X_k\} \cup pa(X_k)$. Por outro lado, pode existir conhecimento pericial sobre as relações entre as variáveis. Esse conhecimento pode ser incorporado no modelo sob forma das referidas probabilidades condicionadas. A vantagem da utilização do grafo bayesiano em alternativa à previsão directa pelo perito das características desconhecidas do objecto situa-se a dois níveis. Em primeiro lugar, o grafo bayesiano suporta a incerteza associada a essas previsões. Em segundo lugar, o grafo bayesiano permite ao perito incorporar o seu conhecimento de relações sobre conjuntos restritos de variáveis e não sobre a totalidade de variáveis envolvidas simultaneamente.

É de notar que o modelo quantitativo de relações entre as variáveis definido através do grafo bayesiano tem $\sum_{i=1}^{m}(n_i \prod_{j:X_j \in pa(X_i)} n_j)$ parâmetros e que o modelo simples, que corresponde à distribuição conjunta da totalidade das variáveis, tem $\prod_{i=1}^{m} n_i$ parâmetros. Quanto menor for o número máximo de pais das variáveis do grafo mais vantajoso será o primeiro modelo relativamente ao segundo. Por exemplo, se os objectos forem descritos por 10 variáveis que podem tomar, cada uma, três modalidades, o número de parâmetros do modelo simples será 59 049. Para a mesma situação e para o caso de cada variável ter dois pais o número de parâmetros do grafo bayesiano será 270.

4 — Definição de uma medida de semelhança

Uma distância é uma aplicação que associa a um par de objectos um número real. Existe, obviamente, uma relação de simetria entre a distância e a medida de semelhança.

Para o agrupamento de objectos descritos por variáveis que tomam um número finito de modalidades Hand (1981) refere um conjunto de distâncias

baseadas no número de características comuns. Essas distâncias diferem pelo facto de em algumas delas serem ponderadas as características ou pelo facto de, no caso de as variáveis serem ordinais, ser tida em consideração a relação de ordem entre as modalidades da variável. No caso mais simples a distância é definida como sendo o número de características distintas entre dois objectos (x_1 e x_2), ou seja $d(x_1, x_2) = \sum_{k=1}^{m} \delta_k(x_1, x_2)$, com $\delta_k(x_1, x_2) = 0$ se $x_k(x_1) = x_k(x_2)$ e $\delta_k(x_1, x_2) = 1$ caso contrário. No caso de as características serem ponderadas a distância tranforma-se em $d(x_1, x_2) = \sum_{k=1}^{m} p_k\delta_k(x_1, x_2)$, sendo $p_1, \ldots, p_m$ os coeficientes de ponderação.

Será proposta uma medida de semelhança para objectos de características parcialmente desconhecidas análoga às medidas atrás consideradas. Essa medida, $s(x_1, x_2)$ será definida como $s(x_1, x_2) = \sum_{k=1}^{m} p_k\delta_k(x_1, x_2)$, sendo agora δ_k tal que s seja uma medida de semelhança, isto é, por forma a que x_1 e x_2 sejam tanto mais semelhantes quanto maior for o valor de $s(x_1, x_2)$. Partindo do pressuposto que os objectos são representados probabilisticamente da forma referida em 2, $\delta_k(x_1, x_2)$ pode ser definido por forma a que tome o valor 1 para dois objectos cuja k-ésima característica seja idêntica e conhecida, isto é, para o caso em que $\exists_j$, $p_{kj}(x_1) = p_{kj}(x_2) = 1$ e tome o valor 0 para o caso em que a k-ésima característica seja distinta e conhecida. Considerar-se-á que a semelhança, relativamente à k-ésima característica, entre um objecto x e qualquer outro objecto é nula se essa característica for completamente desconhecida para x. As situações intermédias (características parcialmente desconhecidas) correspondem a valores para δ_k entre 0 e 1.

Para definir δ_k de acordo com o comportamento geral referido atrás será necessário atender ao grau de conhecimento da k-ésima característica, ao facto dessa variável ser ou não ordinal e, evidentemente, ao facto da característica ser ou não comum aos objectos considerados. Para atender a esses três aspectos considerar-se-á δ_k tem a forma $\delta_k = (D_kO_kC_k)^{\frac{1}{3}}$. O primeiro termo, D_k, medirá a semelhança entre as distribuições de probabilidade dos objectos. O segundo termo, O_k, medirá a dispersão das modalidades da k-ésima variável e envolverá coeficientes que traduzem e relação de ordem existente entre essas modalidades. O terceiro termo, C_k, medirá a incerteza sobre o conhecimento da k-ésima característica dos objectos. Por forma a considerar as possíveis relações de ordem considerar-se-ão coeficientes $(\sigma_{k1}, \ldots, \sigma_{kj}, \ldots \sigma_{kn_k})$ que exprimem, se existir, a semelhança relativa das modalidades da k-ésima variável. Apenas relações de ordem total serão consideradas. Consideram-se as modalidades $(x_{k1}, \ldots, x_{kn_k})$ ordenadas. Pode-se associar a $(x_{k1}, \ldots, x_{knk})$ o vector $(\sigma_{k1}, \ldots, \sigma_{kn_k})$, sendo $\sigma \leq \sigma_{kj} \leq \sigma_{kj+1} \leq 1$ com $\sigma_{k1} = 0$ e $\sigma_{knk} = 1$, tais que $(\sigma_{kj+1}, - \sigma_{kj})$ quantifica a dessemelhança entre x_{kj} e x_{kj}. O centro de gravidade da distribuição da variável X_k, para o objecto x, pode ser representado por $\sigma_k(x)$, sendo $\sigma_k(x) = \sum_{j=1}^{nk} p_{kj}(x)\ \sigma_{kj}$.

$D_k(x_1, x_2)$ pode ser definido, simplesmente, a partir da distância euclidiana entre x_1 e x_2, cada um dos quais representado pelo vector de probabilidades $(p_{k1}, \ldots, p_{knk})$, ou seja:

$$D_k(x_1, x_2) = 1 - \frac{1}{\sqrt{2}} \sqrt{\sum_{j=1}^{nk} (p_{kj}(x_1) - p_{kj}(x_2))^2 \ |\sigma_k(x_1) - \sigma_k(x_2)|}.$$

As constantes presentes na expressão anterior visam apenas transformar a distância numa medida de semelhança e estandardizá-la por forma a tomar valores entre 0 e 1. O termo $|\sigma_k(x_1) - \sigma_k(x_2)|$ permite considerar as relações ordem entre as modalidades de x_k. Este termo anula-se se os centros de gravidade das distribuições de x_1 e de x_2 forem iguais e toma o valor 1 se as variáveis forem conhecidas completamente distintas, isto é, se $p_{k_1}(x_1) = 1$ e $p_{knk}(x_2) = 1$. $D_k(x_1, x_2)$ toma o valor máximo, 1, se as distribuições $p_k(x_1)$ e $p_k(x_2)$ forem idênticas e toma o valor mínimo, 0, se $|\sigma_k(x_1) - \sigma_k(x_2)|$ tomar o valor 1.

$O_k(x_1, x_2)$ será definido como um produto de três termos. Os dois primeiros, $O'_k(x_1)'$ e $O'_k(x_2)$, medirão, respectivamente, a concentração da distribuição $p_k(x_1)$ e da distribuição $p_k(x_2)$. O terceiro termo, $O''_k(x_1, x_2)$, medirá o afastamento entre os centros de gravidade das distribuições dos objectos. Ter-se-á, então, $O_k(x_1, x_2) = O'_k(x_1)\ O'_k(x_2)\ O''_k(x_1, x_2)$, sendo:

$$O'_k(\sigma) = 1 - \sqrt{\sum_{j=1}^{nk} p_{kj}(x)(\sigma_{kj} - \sigma_{k(x)})^2} \text{ e } O''_k(x_1, x_2) = (1 - |\sigma_k(x_1) - \sigma_k(x_2)|)^3$$

O termo sob a raiz é a variância empírica dos σ_{kj}, considerando que a frequência relativa para cada σ_{kj}, é p_{kj}. O expoente do termo O''_k é estabelecido por forma que esse termo pese razoavelmente no valor de δ_k. O'_k tomará o valor máximo, 1, se a k-ésima característica de x for conhecida ou se σ_{kj}, constante, $\forall_j$, situação que será utilizada para traduzir o facto de x_k não ser ordinal. O'_k tomará valores sempre não negativos e será tanto menor quanto mais dispersa for a distribuição $p_k(x)$. É conveniente notar que, pelo facto de se impor que $\sigma_{k1=0}$ e que $\sigma_{knk} = 1$, O'_k é função das diferenças $(\sigma_{kj+1} - \sigma_{kj})$ mas não depende do valor absoluto dos σ_{kj}.

Pode considerar-se, de forma semelhante, que $C_k(x_1, x_2) = C'_k(x_1, x_2)$ sendo $C'_k(x)$ uma medida de certeza sobre a k-ésima característica do objecto x. C_k deverá ter um valor próximo de 1 apenas se ambos os objectos estiverem bem caracterizados. Com base na função de informação pode ser definido C'_k como sendo:

$$C'_k(x) = 1 + \frac{1}{\log n_k} \cdot \sum_{j=1}^{nk} p_{kj}(x) \log p_{kj}(x).$$

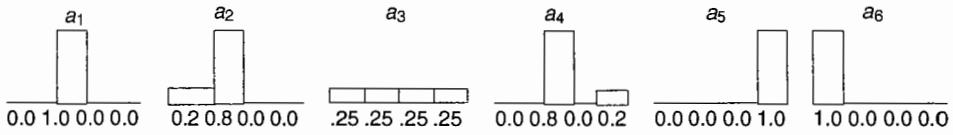
No caso de p_{kj} ser nulo considera-se que o termo $p_{kj(x)} \log p_{kj(x)}$ do somatório da expressão anterior é nulo. O termo $\sum_j p_{kj} \log p_{kj}$, designado entropia, exprime (v. Thornton, 1992) o nível esperado de incerteza sobre a k-ésima característica. Se a k-ésima característica de x for conhecida, então $C'_k(x) = 1$, e se for completamente desconhecida, isto é, se a distribuição $p_k(x) = (\frac{1}{n_k}, \frac{1}{n_k}, \ldots, \frac{1}{n_k})$, então $C'_k(x) = 0$.

Pelo que atrás foi dito tem-se então que

$$s(x_1, x_2) = \sum_{k=1}^{m} p_k \delta_k(x_1, x_2),\ \delta_k(x_1, x_2) = [Dk(x_1, x_2)\ O'_k(x_1)\ O'_k(x_2)\ O''_k(x_1, x_2)\ C'_k(x_1)\ C'_k(x_2)]^{\frac{1}{3}} \text{ e } 0 \leq p_k$$

toma valores entre 0 e $\Sigma \rho_k$. Os coeficientes ρ_k podem ser definidos em função do problema e das variáveis envolvidas, atribuindo maiores valores de ρ_k às variáveis que são consideradas mais importantes. Caso não haja nenhum conhecimento à partida sobre este aspecto os coeficientes de ponderação podem ser unitários.

Para ilustrar as funções atrás definidas considere-se o seguinte exemplo. Sejam a_1, a_2, a_3, a_4, a_5 e a_6 seis objectos. Esses objectos poderiam ser, por exemplo, empresas e a *k*-ésima variável poderia caracterizar o nível de penetração dessas empresas num determinado mercado. Suponha-se que o nível pode ser muito baixo, baixo, médio ou elevado. Observem-se distribuições hipotéticas $\rho_k(x)$ para essa variável ordinal com quatro modalidades e para esses objectos:



Considere-se que o vector de semelhança entre as modalidades de X_k é (0.0, 0.33, 0.67, 1.0). Os valores para as funções atrás definidas podem ser observados na tabela 1.

TABELA 1

Valores obtidos para os termos de δ_k para os objectos a_1, a_2, a_3, a_4, a_5 e a_6

x_1	x_2	$\sigma_k(x_1)$	$\sigma_k(x_2)$	$D_k(x_1, x_2)$	$O'_k(x_1)$	$O'_k(x_2)$	$O''_k(x_1, x_2)$	$C'_k(x_1)$	$C'_k(x_2)$	$\delta_k(x_1, x_2)$
a_1	a_1	0,330	0,330	1,000	1,000	1,000	1,000	1,000	1,000	1,000
a_1	a_2	0,330	0,264	0,987	1,000	0,823	0,815	1,000	0,639	0,752
a_1	a_3	0,330	0,500	0,896	1,000	0,330	0,572	1,000	0,000	0,000
a_1	a_4	0,330	0,464	0,973	1,000	0,640	0,649	1,000	0,639	0,639
a_1	a_5	0,330	1,000	0,330	1,000	1,000	0,036	1,000	1,000	0,229
a_1	a_6	0,330	0,000	0,670	1,000	1,000	0,301	1,000	1,000	0,587
a_5	a_6	1,000	0,000	0,000	1,000	1,000	0,000	1,000	1,000	0,000

Como se pode verificar δ_k é mínimo para (a_5, a_6), que são objectos conhecidos e totalmente distintos relativamente à *k*-ésima variável. δ_k é também nulo para (a_1, a_3) porque a_3 é desconhecido relativamente à *k*-ésima variável. Também se pode verificar que δ_k para (a_1, a_2), é superior do que para (a_1, a_4). O facto de a_1 e a_5 serem conhecidos e distintos relativamente ao nível de penetração no mercado não implica, como se pode observar, que a semelhança entre essas duas empresas seja nula pois a variável é ordinal e pode-se, portanto, aceitar que existe alguma semelhança entre a modalidade «baixa» e a modalidade «elevada», dado que existe uma modalidade «muito baixa», que é, essa sim, totalmente distinta de «elevada».

5 — Descrição do algoritmo para determinação das classes

Estando definida uma medida de semelhança entre dois objectos, a determinação das *N* classes pretendidas é realizada da forma convencional.

Como foi referido em 1 será utilizado um método hierárquico para construir as classes. Os métodos hierárquicos diferem entre si, por um lado, pelo conjunto inicial de classes considerado, podendo ser divisivos ou aglomerativos. Os métodos hierárquicos diferem entre si, por outro lado, pela forma como aplicação S é definida a partir da aplicação s, podendo ser *single-link* ou método do vizinho mais próximo (se $S(x_1, C) = \max_{x_2 \in C} s(x_1, x_2)$), *complete-link* ou método do vizinho mais afastado (se $S(x_1, C) = \min_{x_2 \in C} s(x_1, x_2)$), ou métodos intermédios como, por exemplo, *average link* em que a semelhança entre duas classes é a média das semelhanças entre os elementos dos pares constituídos por um elemento de uma classe e um da outra classe.

Em Anderberg (1973) são referidos 12 métodos hierárquicos entre os quais se encontram os três métodos sumariamente descritos atrás, e que são utilizados para a realização da aplicação apresentada em 6.

Os algoritmos implementados baseiam-se no procedimento geral para classificação não supervisionada hierárquica aglomerativa descrito em Anderberg (1973). Esse procedimento pressupõe a existência de uma medida de semelhança entre dois conjuntos de objectos. A construção dessa medida é trivial a partir da definição da medida de semelhança S anteriormente referida. O procedimento consiste em: 1) construir, inicialmente, N classes constituídas, cada uma, por um objecto; 2) reunir as duas classes mais semelhantes; 3) recalcular as distâncias entre as classes; e 4) voltar ao passo 2 até que o número de classes pretendido seja alcançado.

É frequente ilustrar o processo de classificação graficamente através de um dendrograma. O dendrograma representa de uma forma muito clara as sucessivas reuniões de classes e os sucessivos valores da medida de semelhança entre as classes que são reunidas. A análise do dendrograma é muito útil para a tomada de decisão sobre o número de classes a considerar quando esse número não é fixado à partida. Essa ferramenta de análise não será, no entanto, considerada neste trabalho.

6 — Exemplo de aplicação da metodologia proposta

Os objectos considerados para esta aplicação são os distritos de Portugal Continental e as variáveis consideradas para descrever esses objectos são variáveis climáticas, económicas e ecológicas descritivas do sector florestal de cada distrito. As variáveis utilizadas são de natureza contínua e, consequentemente, foi necessário previamente discretizá-las para poder aplicar a metodologia proposta. A discretização de cada variável foi realizada através de um método de classificação não supervisionada (o método hierárquico aglomerativo do vizinho mais próximo) por forma a identificar os principais intervalos de valores. Cada um dos intervalos obtidos corresponde,

segundo as definições atrás referidas, a uma modalidade de uma variável discreta. As variáveis, as modalidades consideradas e a descrição dessas modalidades são apresentadas na tabela 2:

TABELA 2

Descrição das variáveis e das modalidades consideradas

Variáveis	Descrição	Fonte	Modalidades	Descrição
1 — Prop. sup. flor. ...	Proporção da área territ. ocupada por sup. florestal.	(*)	1) Baixa	< 15 %
			2) Média	15–25 %
			3) Elevada	25–33,3 %
			4) Muito elevada	> 33,3 %
2 — Clima	Temperatura média (°C)/ precipitação total (mm).	(***)	1) Seco e quente	> 3
			2) Médio	1,5–3
			3) Húmido e fresco ...	< 1,5
3 — Coberto flor	Proporção de área florestal ocupada por sobro e azinho.	(**)	1) Elevada	> 75 %
			2) Média	22,5–75 %
			3) Baixa	< 22,5 %
4 — Incêndios	Área de povoamento ardida/área florestal.	(**)	1) Baixo	< 3 %
			2) Elevado	> 3 %
5 — Seguros	Valores segurados (escudos)/área florestal (ha).	(**)	1) Baixo	< 120
			2) Médio	120–1000
			3) Elevado	> 1000
6 — Caminhos	Caminhos florestais construídos (m)/área florestal (ha).	(**)	1) Baixo	< 10
		(**)	2) Elevado	> 10
7 — Reg. flor.	Proporção da área florestal sujeita ao regime florestal.	(**)	1) Baixa	< 45 %
			2) Elevada	> 45 %
8 — Arboriz. DGF	Área arborizada pela Direcção-Geral das Florestal.	(*)	1) Baixa	< 1500 ha
			2) Elevada	> 1500 ha

(*) *Estatísticas Agrícolas de 1988*, INE.
(**) *Estatísticas Agrícolas de 1988*, INE (adaptado).
(***) «Atlas de Portugal», *Selecções do Reader's Digest*, 1988 (adaptado).

Os objectos podem ser representados de forma sintética através de oito dígitos, cada qual correspondente a uma variável. O valor do *k*-ésimo dígito representa a modalidade que é tomada pela *k*-ésima variável. Por exemplo o objecto 43312111 (que corresponde ao distrito de Aveiro) tem proporção de superfície florestal elevada (4), tem clima húmido e fresco (3), e assim sucessivamente. No caso de uma característica ser desconhecida pode-se utilizar uma representação semelhante para os objectos. Pode-se atribuir o valor 0 ao dígito correspondente a essa característica. Por exemplo, o distrito de Aveiro pode ser representado por 40312111 se o clima for desconhecido.

As características dos 18 distritos de Portugal Continental são, de acordo com a representação atrás descrita, apresentadas na tabela 3:

TABELA 3

Descrição dos objectos

Aveiro	43312111	Beja	31112121	Braga	43311212
Bragança	12211212	Castelo Branco	43311111	Coimbra	43322111
Évora	42112121	Faro	21212111	Guarda	23322111
Leiria	42321111	Lisboa	12323121	Portalegre	42113121
Porto	43311111	Santarém	42212121	Setúbal	42212121
Viana	43321221	Vila Real	33322222	Viseu	43322111

Os dados serão processados da seguinte forma. Os objectos de características completamente conhecidas serão agrupados em classes através de três algoritmos de classificação supervisada hierárquicos e aglomerativos: vizinho mais próximo, vizinho mais afastado e semelhança média. Considerar-se-ão para o número de classes os valores $N=2$, $N=3$, $N=4$ e $N=5$.

Por forma a testar o método de classificação para objectos com características parcialmente desconhecidas apagar-se-á, primeiramente, um determinado número de características. Por apagamento entende-se a perda do conhecimento da modalidade da variável, ou seja, e em termos de descrição probabilística do objecto, a substituição de uma distribuição, respeitante à k-ésima variável, com a forma (0,...,0,1,0,...,0) por uma distribuição com a forma $\left(\frac{1}{nk},\ldots\frac{1}{nk}\right)$.

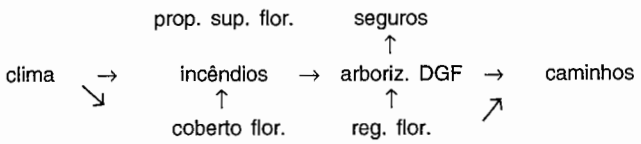
O procedimento adoptado é o que se descreve em seguida. Apagam-se, aleatoriamente, 5 %, 10 %, 15 % e 20 % das características dos objectos e aplica-se o método de classificação proposto aos objectos obtidos dessa forma, utilizando-se os três algoritmos atrás referidos. Estes objectos são também classificados através dos mesmos três algoritmos, sem serem sujeitos ao procedimento de complementação das características. Obtêm-se assim, para cada algoritmo e para cada valor de N considerado, classificações correspondentes aos objectos completamente conhecidos, aos objectos com 5 %, 10 %, 15 % e 20 % de características desconhecidas e aos objectos com 5 %, 10 %, 15 % e 20 % de características desconhecidas mas estimadas pelo método descrito neste trabalho. Essas classificações são realizadas para 10 escolhas aleatórias de características a apagar. É de notar que são apagadas x% das 18×8 características e não x% das 8 características de cada objecto.

Para poder aplicar o método proposto é necessário definir o grafo bayesiano que estabelece relações entre as variáveis envolvidas.

A componente gráfica é, neste exemplo, construída de forma unicamente pericial. Essa componente é apresentada na figura 1. Como se pode observar, uma das variáveis, a proporção de área florestal, foi considerada independente das outras. Este facto tem como consequência que, se essa variável for desconhecida para um objecto, a estimação da sua distribuição não é baseada nas outras características (conhecidas) do objecto.

FIGURA 1

Componente gráfica do grafo bayesiano



A componente probalística é obtida a partir dos dados e do conhecimento pericial sobre as relações entre as variáveis. Como se trata de uma aplicação cujo objectivo é testar o desempenho médio do método de classificação proposto e para tal é repetido o algoritmo de classificação um elevado número de vezes, estabeleceu-se uma forma automática de tomar em consideração o conhecimento pericial e a informação dos dados para a definição da componente probabilística, ou seja, para a escolha das probabilidades condicionadas que a constituem.

Supôs-se, então, que o perito se baseia, parcialmente, na informação das características conhecidas dos objectos para a escolha das probabilidades e que a decisão é tomada da forma seguinte. No caso de a probabilidade, estimada com base nos dados, de a variável X_k tomar a modalidade x_{kj}, dado um determinado estado dos pais de X_k, ser consistente com o conhecimento do perito este aceita essa estimativa como parâmetro do grafo bayesiano. No caso de essa estimativa ser inconsistente com o conhecimento do perito este indica um novo valor que passa a ser o parâmetro do grafo bayesiano.

Para simular o comportamento atrás descrito estima-se, inicialmente e a partir dos dados, os parâmetros da componente probabilística do grafo bayesiano (as probabilidades condicionadas respeitantes às famílias definidas pelo perito na componente gráfica), obtendo dessa maneira um modelo de referência para a componente probalística. Finalmente, no decorrer de cada execução do algoritmo de classificação, utilizam-se as probabilidades desse modelo de referência se os valores estimados a partir dos dados se distanciarem mais de 0,1 dos valores do modelo de referência e utilizam-se as estimativas obtidas a partir das características conhecidas e da componente gráfica caso contrário.

Na tabela 4 apresentam-se os 58 parâmetros da componente probabilística que constituem o modelo de referência. Nessa tabela os estados das famílias são representados da forma sintética descrita no início de 6. As variáveis que não pertencem à família tomam, na tabela 4, o valor 0.

TABELA 4

Componente probabilística do grafo bayesiano

k	Estado de X_k, $pa(X_k)$	Estado de $P(X_k/pa(x_k))$	k	Estado de X_k, $pa(X_k)$	Estado de $P(X_k/pa(x_k))$	k	Estado de X_k, $pa(X_k)$	Estado de $P(X_k/pa(x_k))$
1	10000000	0,111	2	1000000	0,111	4	1110000	1,000
1	20000000	0,111	2	2000000	0,389	4	1120000	0,000
1	30000000	0,111	2	3000000	0,500	4	1210000	1,000

k	Estado de X_k, $pa(X_k)$	Estado de $P(X_k/pa(x_k))$	k	Estado de X_k, $pa(X_k)$	Estado de $P(X_k/pa(x_k))$	k	Estado de X_k, $pa(X_k)$	Estado de $P(X_k/pa(x_k))$
1	40000000	0,666	3	1100000	0,500	4	1220000	0,000
5	1001	0,267	3	1200000	0,500	4	1310000	0,500
5	2001	0,600	3	1300000	0,000	4	1320000	0,500
5	3001	0,133	3	2100000	0,285	4	2110000	1,000
5	1002	0,653	3	2200000	0,428	4	2120000	0,000
5	2002	0,326	3	2300000	0,285	4	2210000	1,000
5	3002	0,000	3	3100000	0,000	4	2220000	0,000
7	10	0,555	3	3200000	0,000	4	2310000	0,000
7	20	0,444	3	3300000	1,000	4	2320000	1,000
8	10011	0,667	6	111	1,000	4	3110000	0,500
8	10012	0,333	6	211	0,000	4	3120000	0,500
8	10021	1,000	6	112	0,000	4	3210000	0,500
8	10022	0,000	6	212	1,000	4	3220000	0,500
8	20011	1,000	6	121	0,857	4	3310000	0,444
8	20012	0,000	6	221	0,143	4	3320000	0,555
8	20021	0,667	6	122	0,000			
8	20022	0,333	6	222	1,000			

Os valores dos ρ_k e dos σ_{kj} que se apresentam na tabela 5 completam o conjunto de dados necessários para a execução dos diversos algoritmos de classificação. Estes valores foram escolhidos pericialmente em função da natureza das variáveis utilizadas. Note-se que todas as variáveis são ordinais.

TABELA 5

Coeficientes da medida de semelhança entre objectos

k	ρ_k	σ_{kj}			
1	1,5	0	0,33	0,67	1
2	1	0	0,5	1	
3	0,8	0	0,5	1	
4	0,8	0	1		
5	0,8	0	0,5	1	
6	0,8	0	1		
7	1	0	1		
8	0,8	0	1		

Para analisar o desempenho do método proposto é necessário comparar o resultado da classificação dos objectos de características completamente conhecidas relativamente aos resultados do método proposto e relativamente ao resultado da classificação dos objectos de características parcialmente desconhecidas e não complementadas de forma alguma. A comparação processa-se da seguinte forma. Dado um determinado número de classes, afectam-se as classes obtidas para os objectos de características completamente conhecidas às classes obtidas para os objectos de características parcialmente desconhecidas maximizando o número de elementos pertencentes às intersecções entre as classes afectadas entre si. A medida de comparação é esse número máximo de elementos. Por exemplo, se a classificação de quatro objectos $\{a_1, a_2, a_3, a_4\}$ com características conhecidas originar as classes $\{a_1, a_4\}$ e $\{a_2, a_3\}$

e se a classificação dos mesmos quatro objectos com características desconhecidas originar as classes $\{a_1, a_2, a_4\}$ e $\{a_3\}$ a medida de comparação desses dois resultados valerá 3 pois a_1, a_4, e a_3 pertencem às intersecções $\{a_1, a_4\} \cap \{a_1, a_2, a_4\}$ e $\{a_2, a_3\} \cap \{a_3\}$. A outra afectação possível das classes ($\{a_1, a_4\}$ com $\{a_3\}$, e $\{a_2, a_3\}$ com $\{a_1, a_2, a_4\}$) teria, nas intersecções, apenas um objecto, a_2.

Pode-se ilustrar também esta medida com os dados relativos aos distritos. Considerando $N = 5$ e o método de classificação do vizinho menos semelhante as classes obtidas para os objectos de características completamente conhecidas (v. tabela 3) são: { Aveiro, Castelo Branco, Porto, Coimbra, Viseu, Leiria, Guarda, Braga}, { Beja, Évora, Portalegre, Santarém, Setúbal, Faro }, { Bragança }, { Lisboa }, e { Viana, Vila Real }. As classes obtidas, utilizando também o método do vizinho menos semelhante, para os objectos relativamente aos quais foram apagadas 10 % das características e relativamente aos quais foi aplicado o método proposto neste trabalho para estimação das características desconhecidas são: { Aveiro, Santarém, Setúbal, Castelo Branco, Porto, Leiria, Coimbra, Viseu }, { Beja, Faro, Évora, Portalegre }, { Braga, Viana, Vila Real }, { Bragança }, e { Guarda, Lisboa }. A medida de comparação das duas classificações é 14.

Na tabela 6 são apresentadas as médias dos valores obtidos para essas medidas para as 10 amostras aleatórias consideradas. Note-se que o valor máximo que pode ser obtido na comparação é 18, o número de objectos a classificar. Analisando a tabela 6 verifica-se que os resultados obtidos com o método proposto (coluna «Compl.») são, na quase totalidade das situações, superiores aos resultados obtidos quando as características desconhecidas não são estimadas (coluna «Não c.») As médias de todos os valores da medida de comparação são, respectivamente, 14,90625 e 13,82292.

TABELA 6

Comparação dos resultados

Percentagem de características apagadas	Número de classes	*Single-link*		*Complete-link*		*Average-link*	
		Compl.	Não c.	Compl.	Não c.	Compl.	Não c.
5	5	15,500	15,300	16,500	14,800	16,600	15,500
	4	15,100	14,900	16,700	14,600	15,100	13,700
	3	16,800	16,200	14,400	12,000	13,200	12,400
	2	17,400	16,800	14,200	12,700	16,000	16,600
	média	16,200	15,800	15,450	13,525	15,225	14,550
10	5	14,700	14,500	15,000	13,700	16,500	13,100
	4	14,700	14,100	14,700	13,600	14,300	13,300
	3	16,800	15,400	13,100	11,000	11,400	10,900
	2	16,500	15,900	13,800	13,700	17,200	17,100
	média	15,675	14,975	14,150	13,000	14,850	13,600
15	5	14,100	14,200	14,700	13,200	16,000	13,600
	4	14,000	14,300	14,100	11,900	12,800	13,400
	3	15,800	15,300	12,900	10,500	10,900	11,100
	2	16,300	15,900	15,100	12,700	17,000	16,500
	média	15,050	14,925	14,200	12,075	14,175	13,650

Percentagem de características apagadas	Número de classes	*Single-link*		*Complete-link*		*Average-link*	
		Compl.	Não c.	Compl.	Não c.	Compl.	Não c.
20	5	14,200	13,800	14,900	12,800	15,400	12,700
	4	14,200	14,000	14,600	12,600	14,400	12,000
	3	15,500	15,100	12,800	10,400	12,700	12,100
	2	16,400	15,800	14,300	11,500	16,200	16,300
	média	15,075	14,675	14,150	11,825	14,675	13,275
média	—	15,500	15,094	14,488	12,606	14,731	13,769

7 — Conclusões

Em classificação não supervisada o objectivo é identificar um conjunto de classes ou grupos que se ajustem aos dados. O problema é referido como o de procurar «grupos naturais» para os objectos (v. Anderberg, 1973). É a escolha das variáveis e da medida de semelhança que dá um sentido operacional ao termo «grupos naturais».

Neste trabalho a questão da escolha da medida de semelhança entre grupos não foi desenvolvida. As medidas utilizadas e descritas em 5 são medidas clássicas e muito testadas. Supõe-se que qualquer dos métodos de classificação utilizados revela uma estrutura de classes própria ao conjunto de objectos, não se pondo em causa a qualidade dessa classificação.

A questão que aqui é analisada é a da aplicação dos métodos de classificação convencionais a objectos de características parcialmente desconhecidas. Tomando como referência as classes reveladas por cada um dos algoritmos de classificação, pretendeu-se desenvolver um método que permitisse «recuperar» essa estrutura no caso de serem desconhecidas características dos objectos. Para atingir esse objectivo propôs-se uma metodologia que permite estruturar e explorar a informação das características conhecidas dos objectos e o conhecimento pericial existente sobre o domínio do problema.

Com base nos resultados apresentados na tabela 6 pode concluir-se que o método proposto, para a aplicação realizada, é eficaz, no sentido de completar a informação desconhecida, pois origina resultados mais próximos dos resultados obtidos em condições de conhecimento total das características dos objectos do que a classificação apenas baseada nas características conhecidas. Verifica-se que se obtêm classes semelhantes às classes de referência mesmo quando a proporção de características apagadas atinge 20 %. Este resultado parece dever-se menos ao facto de as classes obtidas serem estáveis do que ao facto do método ter um bom desempenho pelas duas razões seguintes. Por um lado, como se observa na tabela 6, a medida de comparação degrada-se, para o caso da não complementação dos objectos, à medida que a proporção de objectos apagados aumenta mas essa tendência é muito menos evidente para o caso dos objectos cujas características desconhecidas foram estimadas através do método proposto. Por outro lado, a utilização do grafo bayesiano permite realmente obter algum conhecimento sobre as características desconhecidas. Relativamente a este último aspecto observe-se a tabela 7 na qual são apresentadas, para os seis distritos de menor ordem alfabética, as

distribuições das modalidades das variáveis 2 a 7, para o caso em que as características são completamente conhecidas (linha superior) e para o caso em que 20 % da totalidade das características (dos 18 distritos) são apagadas aleatoriamente e as respectivas distribuições são estimadas através do grafo bayesiano (linha inferior). As variáveis 1 e 8 foram excluídas da tabela 7 porque nenhuma delas correspondia a características apagadas nos seis distritos considerados (tal como acontece, aliás, para a variável 5).

TABELA 7

Distribuições originais e estimadas através do grafo bayesiano após apagamento de 20 % das características

Distritos	Variáveis e respectivas modalidades														
	2			3			4		5			6		7	
	1	2	3	1	2	3	1	2	1	2	3	1	2	1	2
Aveiro			1,00			1,00	1,00			1,00		1,00		1,00	
			1,00			1,00	0,42	0,57		1,00		0,92	0,07	0,61	0,38
Beja	1,00			1,00			1,00			1,00		1,00			1,00
	1,00			1,00			1,00			1,00		0,91	0,08	0,51	0,48
Braga			1,00			1,00	1,00		1,00				1,00	1,00	
	0,21	0,52	0,25	0,36	0,36	0,26	0,96	0,03	1,00				1,00	1,00	
Bragança		1,00			1,00		1,00		1,00				1,00	1,00	
		1,00			1,00		1,00		1,00				1,00	1,00	
Castelo Branco			1,00			1,00	1,00		1,00			1,00		1,00	
			1,00			1,00	1,00		1,00			1,00		0,54	0,45
Coimbra			1,00			1,00		1,00		1,00		1,00		1,00	
			1,00			1,00		1,00		1,00		0,96	0,03	1,00	

Note-se que em alguns casos a distribuição obtida é mais afastada da distribuição real (característica conhecida) do que a distribuição uniforme $p_k = \left(\frac{1}{nk}, \ldots, \frac{1}{nk}\right)$. Na maior parte das situações, no entanto, verifica-se o contrário. Esse comportamento justifica os resultados apresentados na tabela 6 porque mostra que o método proposto permite «recuperar», parcialmente, as características desconhecidas dos objectos.

Um aspecto que deve ser comentado é o da intervenção do perito e das possíveis implicações dessa intervenção no resultado da classificação.

O perito intervém na construção da componente gráfica do grafo bayesiano, estabelecendo as relações que considera pertinentes entre as variáveis. Essas especificações reflectem-se na componente probabilística, que é constituída por um conjunto de probabilidades condicionadas respeitantes às famílias definidas na componente gráfica. Mesmo que essas probabilidades sejam estimadas a partir dos dados, a «leitura» da informação contida nesses dados é feita pressupondo o modelo qualitativo de relações de dependência entre as variáveis, ou seja, a componente gráfica. Este aspecto é fundamental porque a construção da componente gráfica, que pode ser realizada por qualquer perito do domínio do problema de uma forma extremamente simples, permite que seja

tirado partido da informação contida nos dados de modo muito mais eficiente. Na verdade, ao serem agrupadas as variáveis em famílias que constituem os «processadores» do método de propagação de evidências, são criadas condições para que a estimação dos parâmetros probabilísticos que descrevem as relações entre as variáveis se baseiem num maior número de observações. Por exemplo, na aplicação apresentada em 6, os 18 objectos permitiram estimações não nulas de 44 dos 58 parâmetros da componente probabilística.

O perito pode, para além disso, intervir directamente na componente probabilística, especificando parâmetros e completando também dessa forma o conhecimento sobre o domínio do problema e pode, finalmente, especificar os coeficientes da medida de semelhança (v. tabela 5).

Uma questão que também ocupa um lugar importante neste trabalho é da definição da medida de semelhança. Considerou-se importante que essa medida atendesse a três aspectos: valor das características, ordem das modalidades e conhecimento existente sobre as características. A medida proposta exprime todos esses aspectos mas tem uma forma que poderia, talvez vantajosamente, ser simplificada. Em particular, a definição dos parâmetros dessa medida que não dependem dos dados não é justificada por nenhuma razão para além do facto desses parâmetros originarem medidas «aceitáveis» para os objectos com os quais foram testados.

REFERÊNCIAS

ANDERBERG, M.R. (1973), *Cluster Analysis for Applications,* Academic Press, New York.

CAMPAGNOLO, M.L (1992), «Proposta de um método para integração de conhecimento em classificação», dissertação para obtenção do grau de Mestre em Matemática Aplicada à Economia e Gestão, Instituto Superior de Economia e Gestão, Lisboa, não publicado.

GEIGER, D., VERMA, T. e PEARL, J. (1990), *Identifying independence in bayesian networks,* Networks, vol. 20, pp. 507-534.

GNANADESIKAN, R., BLASHFIELD, R.K., BREIMAN, L., DUNE O.J., FRIEDMAN, J.H., KING-SUN FU, HARTIGAN, J.A., KETTENRING, J.R., LACHENBRUCH, P.A., OLSHEN, R.A. and ROHOLF, F.J. (1989), «Discriminant analysis and clustering», *Statistical Science,* vol. 4, n.º 1, pp. 34-69.

HAND, D.J., (1981), *Discrimination and Classification,* John Wiley and Sons, New York.

PEARL, J. (1986), «Fusion, propagation and structuring in belief networks», *Artificial Intelligence,* 29, pp. 241-88.

—— (1987), «Evidencial reasoning using stochastic simulation», *Artificial Intelligence,* 32, pp. 245--57.

SPIEGHELHALTER, D.J., DAWID, A.P., LAURITZEN, S.L., e COWELL, R.G. (1992), «Bayesian analysis in expert systems», *Research report 92-6,* MRC Biostatiscs Unit., Cambridge.

THORNTON, C.J. (1992), *Techniques in Computacional Learning,* Chapman & Hall, London.

WHITTAKER, J. (1990), *Graphical Models in Applied Multivariate Statistics,* Wiley, Chicester.