Sobre a Comparação de Estruturas de Classificação: Coeficientes e suas Distribuições

Fernanda Sousa[1] · Jorge Tendeiro[2]

Resumo Numa Classificação Hierárquica Ascendente a função de comparação entre pares de elementos e o critério de agregação induzem relações estruturais entre os elementos do conjunto a classificar, aqui designadas por estruturas de classificação. Neste trabalho são invocadas razões para a necessidade de comparar pares destas estruturas e são introduzidos coeficientes adequados para tal comparação. As distribuições teóricas assintóticas desses coeficientes são apresentadas e é discutida a sua adequação ao problema em análise. A solução proposta passa pela dedução de distribuições empíricas, por recurso à simulação, e é ilustrada para o caso da comparação de dendrogramas sobre o mesmo conjunto de elementos, recorrendo a métodos de geração aleatória de dendrogramas.

Palavras-chave: Classificação Hierárquica, Dendrograma, Geração Aleatória de Dendrogramas, Coeficientes Ordinais de Comparação, Distribuição de Coeficientes.

1 Introdução

A Análise Classificatória tem por objectivo agrupar um conjunto de objectos num número relativamente pequeno de classes, satisfazendo a condição de que objectos de uma mesma classe sejam mais semelhantes entre si que objectos pertencentes a classes distintas. O trabalho aqui apresentado foca-se nos métodos de Classificação Hierárquica Ascendente (C.H.A.). Da aplicação de um método de

[1]Faculdade de Engenharia e CITTA, Universidade do Porto, fcsousa@fe.up.pt

[2]Departamento de Psicometria e Estatística, Faculdade de Psicologia, Universidade de Groningen, j.n.tendeiro@rug.nl

C.H.A. a um quadro de dados surgem várias entidades que reflectem relações de proximidade entre os elementos a classificar e informação da sua pertença a um mesmo grupo. Entre essas entidades, aqui designadas por estruturas de classificação, salientam-se as produzidas pela função de comparação entre pares de elementos e as associadas ao *output*, sendo os dendrogramas as mais usadas. É bem conhecido e aceite pela comunidade científica (Gordon, 1999) que diferentes opções na aplicação de uma C.H.A. conduzem frequentemente a resultados diferentes, não havendo escolhas reconhecidas como genericamente melhores. É atitude corrente considerar várias opções na aplicação de uma C.H.A. a um dado quadro de dados, daí resultando a necessidade de avaliar o grau de concordância das várias estruturas de classificação produzidas.

A comparação de estruturas de classificação, obtidas para o mesmo conjunto de elementos, é o tema deste trabalho. Serão apresentados coeficientes adequados a este tipo de comparação, bem como uma análise das suas distribuições.

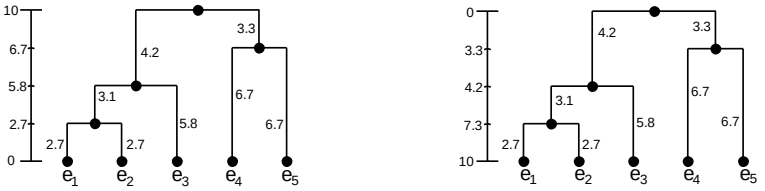
Na Secção 2 são introduzidas definições de apoio ao desenvolvimento do texto. A Secção 3 é dedicada à abordagem proposta para comparar estruturas de classificação. A Secção 4 foca-se nas distribuições assintóticas e empíricas dos coeficientes introduzidos na secção anterior. Uma apresentação e discussão dos resultados são feitas na Secção 5. A Secção 6 é dedicada a algumas conclusões.

2 Classificação Hierárquica Ascendente

Os métodos de C.H.A. têm por base a definição de duas funções de comparação: a **função de comparação entre elementos**, γ, e a **função de comparação entre classes** (critério de agregação), Γ. Dado um conjunto $E=\{e_1, e_2, ..., e_m\}$ de objectos a classificar, a função $\gamma: E \times E \rightarrow \mathbb{R}_0^+$, que pode ser do tipo dissemelhança ou semelhança, mede o grau de parecença entre pares de elementos. A função $\Gamma: P(E) \times P(E) \rightarrow \mathbb{R}_0^+$, onde $P(E)$ é o conjunto das partes de E, mede o grau de parecença entre pares de partes de E.

Como resultado de uma C.H.A. tem-se uma **hierarquia indiciada**, um **dendrograma** ou uma **matriz ultramétrica** sobre E. Uma hierarquia indiciada ou indexada é um par (H,h). H é uma hierarquia, isto é, uma sucessão de partições encaixadas de E. A função h associa a cada parte de E, seja A, o valor da função de comparação entre classes que deu origem ao nível em que A é formada. Um dendrograma é uma árvore ponderada com raiz, em que os nós terminais ou folhas são etiquetados e estão todos à mesma distância da raiz. Os nós internos de um dendrograma podem ser ordenados de acordo com a sua distância relativa às folhas, quando o critério de agregação é do tipo dissemelhança (Figura 1- (a)), ou à raiz, quando o critério de agregação é do tipo semelhança (Figura 1 - (b)). Os valores associados aos nós designam-se por índices de nível e são dados pela função h introduzida acima.

Muitas vezes, mais do que trabalhar com índices de nível, trabalha-se com níveis de agregação (ou níveis de fusão), que são os valores ordinais dos índices de nível. Desta forma, dá-se mais relevância à posição relativa dos nós, em detrimento dos valores reais das distâncias entre eles. Sem prejuízo de generalização, neste trabalho consideram-se dendrogramas completamente binários.



(a) (b)

Figura 1 – (a) Distância relativa às folhas (b) Distância relativa à raiz.

Um dendrograma fica completamente definido por três características:

- **topologia**: diz respeito à forma, ou seja, ignora as etiquetas e os pesos atribuídos aos diferentes ramos; sob este ponto de vista, duas árvores são distintas se possuírem sistemas de bifurcação distintos (Figura 2);
- **etiquetas das folhas**: fixada uma certa topologia, há usualmente diferentes formas de etiquetar as folhas (Figura 3);
- **níveis de agregação**: dois dendrogramas que partilhem a topologia e a etiquetagem podem diferir nos níveis de agregação (Figura 4).

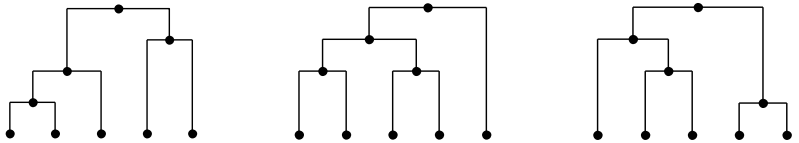
O número de dendrogramas distintos é uma função $d(m)$ do número m de nós terminais; admitindo que não há empates nos níveis de fusão, mostra-se (por exemplo, em Podani, 2000, pg. 126) que

$$d(m) = \frac{m!(m-1)!}{2^{m-1}} \quad (1)$$

Dois dendrogramas que têm a mesma topologia, etiquetagem das folhas e níveis de fusão dizem-se **isomorfos**. Uma representação alternativa e equivalente ao dendrograma é a **matriz ultramétrica** (Figura 5). A cada par de nós terminais $e_i, e_j \in E$ associa-se o valor $h(\{e_i, e_j\})$. Trata-se de uma matriz simétrica com $M = \binom{m}{2}$ entradas relevantes, que representam todas as combinações possíveis de pares de vértices distintos. O nome desta matriz vem do facto dos seus elementos verificarem a **propriedade ultramétrica**, ou seja,

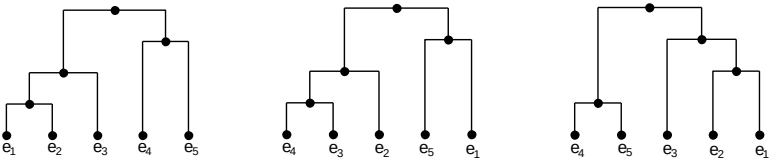
- $h(e_i, e_j) \geq min\{h(e_i, e_k), h(e_k, e_j)\}, \forall e_i, e_j, e_k \in E$, se a função de comparação entre elementos for do tipo semelhança.

- $h(e_i, e_j) \leq max\{h(e_i, e_k), h(e_k, e_j)\}, \forall e_i, e_j, e_k \in E$, se a função de comparação entre elementos for do tipo dissemelhança.



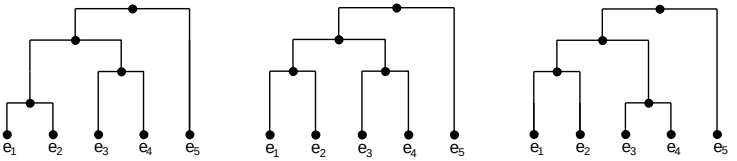
(a) (b) (c)

Figura 2 – As árvores (a) e (b) não são topologicamente idênticas, mas as árvores (a) e (c) são topologicamente idênticas.



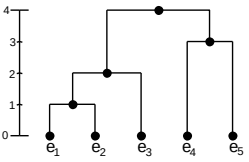
(a) (b) (c)

Figura 3 – As árvores (a), (b), e (c) são todas topologicamente idênticas, mas apenas as árvores (a) e (c) são idênticas quando se consideram as etiquetas.



(a) (b) (c)

Figura 4 – Os dendrogramas (a), (b) e (c) são todos distintos, embora tenham a mesma topologia e etiquetagem.



(a)

$$\begin{array}{c} \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{array} \begin{array}{c} \begin{array}{cccc} e_1 & e_2 & e_3 & e_4 \end{array} \\ \begin{bmatrix} 1 & & & \\ 2 & 2 & & \\ 4 & 4 & 4 & \\ 4 & 4 & 4 & 3 \end{bmatrix} \end{array}$$

(b)

Figura 5 – Dendrograma e respectiva matriz ultramétrica.

Uma **relação binária** sobre um conjunto E é um conjunto $R \subseteq E \times E$. Uma relação binária reflexiva e transitiva diz-se uma **preordem**, enquanto uma relação binária reflexiva, antisimétrica e transitiva diz-se uma **ordem**. Uma relação em que quaisquer dois elementos são comparáveis diz-se **total**.

Seja O uma relação binária sobre o conjunto $E \times E$ que verifica as seguintes condições: (i) $(e_i, e_i) O (e_j, e_k), \forall e_i, e_j, e_k \in E$ (ii) $(e_i, e_j) O (e_j, e_i), \forall e_i, e_j \in E$ e (iii) $(e_j, e_k) O (e_i, e_i) \Longrightarrow e_j = e_k, \forall e_i, e_j, e_k \in E$. Sejam ainda os conjuntos F e G definidos por $F = \left\{\{e_i, e_j\} : e_i, e_j \in E, e_i \neq e_j\right\}$ e $G = \left\{(e_i, e_j) : e_i, e_j \in E, e_i \neq e_j\right\}$.

Diz-se que O é uma **ordenação** (**preordenação**) de E se se verificar uma das três condições (equivalentes) seguintes: (i) O é ordem (preordem) em $E \times E$ (ii) O é ordem (preordem) em F (iii) O é ordem (preordem) total em G. Na Subsecção 3.3 será apresentado um exemplo que ilustra estas noções.

3 Comparação de estruturas de classificação

3.1 Abordagem proposta

O conjunto de valores da função de comparação entre elementos, $\gamma(e_i, e_j)$ com $e_i, e_j \in E$, a matriz ultramétrica, as sucessivas partições da hierarquia, ou o dendrograma revelam, sobre os elementos de E, relações que definem **estruturas de classificação**. À função de comparação sobre pares de elementos associa-se, em geral, uma ordenação de E, enquanto à matriz ultramétrica, a uma partição, ou a um dendrograma se associam preordenações. Escolhas diferentes de funções de comparação (entre pares de elementos ou associadas aos critérios de agregação) produzem, frequentemente, diferentes estruturas de classificação (dendrogramas, matrizes ultramétricas, hierarquias, partições). Para responder, entre outras, a questões como qual a melhor escolha a fazer para as funções de comparação, ou se o resultado da C.H.A. está de acordo com a estrutura inicial dos dados, ou qual o grau de concordância entre os resultados obtidos por dois métodos de classificação sobre um mesmo conjunto de dados, recorre-se à comparação de estruturas de classificação (Nicolau, 1984; Sousa & Nicolau, 2001). Estas preocupações inserem-se na área da Validação em Classificação e têm sido objecto de estudo nas últimas décadas (ver por exemplo Gordon, 1996; Bock, 1996 e Sousa, 2000; Halkidi *et al.*, 2001). Conforme foi já referido anteriormente, este trabalho não responde directamente a estas questões e desenvolve-se a dois níveis.

Num nível mais geral propõe-se uma abordagem ordinal para tratar o problema da comparação de estruturas de classificação associadas a uma C.H.A.. O processo consiste essencialmente em três passos:

1. Associar as estruturas de classificação a preordenações: de facto, é possível fazer corresponder dendrogramas, matrizes ultramétricas e hierarquias indiciadas a preordenações; sendo assim, o nosso propósito — comparar estruturas de classificação — é conseguido através da comparação das preordenações correspondentes.
2. Introduzir e estudar coeficientes de correlação ordinal aplicados a preordenações.
3. Deduzir distribuições adequadas para os coeficientes, tendo em vista a atribuição de significado estatístico aos seus valores.

A um segundo nível neste trabalho aplica-se esta abordagem ao caso particular de dendrogramas obtidos, a partir de um mesmo conjunto de dados, por aplicação de C.H.A. com diferentes escolhas de funções de comparação.

3.2 Coeficientes de comparação de preordenações

Sejam m o número de elementos a classificar e $M = \binom{m}{2} = card(F)$ (notação introduzida na Secção 2). Consideremos duas preordenações ω_1 e ω_2, sobre um mesmo conjunto E, de comprimento M, do tipo (M_1, M_2,…, M_k) e (N_1, N_2,…, N_h). Isto significa que, por exemplo, ω_1 é uma preordenação em que ocorrem k valores distintos, com M_1 elementos de F associados ao primeiro patamar (considerados iguais), M_2 elementos associados ao segundo patamar, etc.. Tem-se portanto $M_1 + M_2 +\ldots+ M_k = M = N_1 + N_2 +\ldots+ N_h$. Definam-se as variáveis aleatórias U_l = "número de ordem do l-ésimo elemento de ω_1" e V_l = "número de ordem do l-ésimo elemento de ω_2", com $l = 1,2,\ldots M$. Aos elementos de um mesmo patamar de uma preordem atribui-se um valor que é a média das ordens que esses elementos teriam se os seus valores fossem diferentes mas consecutivos. Isto permite que a soma dos M valores associados aos elementos de F seja igual a $1 + 2 + \cdots + M = \frac{(1+M)M}{2}$, precisamente o mesmo valor que se obteria se os valores associados aos elementos de F fossem distintos.

Considerem-se as variáveis aleatórias A_{ij}= sgn(U_j-U_i) sgn(V_j-V_i), $1\leq i< j\leq M$, onde sgn representa a função sinal. As variáveis A_{ij} podem tomar três valores:

A_{ij} = 1, se os pares (U_i,V_i) e (U_j,V_j) são **concordantes**;
A_{ij} = -1, se os pares (U_i,V_i) e (U_j,V_j) são **discordantes**;
A_{ij} = 0, se nos pares (U_i,V_i) e (U_j,V_j) ocorre **empate**.

Definam-se ainda as variáveis aleatórias C e D. C denota o número de concordâncias entre as variáveis U e V, ou seja, o número de vezes que A_{ij} = 1 (com $1\leq i< j\leq M$). D denota o número de discordâncias entre U e V, ou seja, o número de vezes que A_{ij} = -1.

Existem na literatura vários coeficientes que podem ser usados para comparar preordenações. Neste estudo, consideram-se os coeficientes de correlação ordinal de Spearman, Kendall e Goodman-Kruskal (Kendall, 1970).

O coeficiente de correlação ordinal de Spearman para as duas preordenações ω_1 e ω_2 é dado por

$$R_S = \frac{\frac{M^3-M}{6} - S(d^2) - T_1 - T_2}{\sqrt{\frac{M^3-M}{6} - 2T_1}\sqrt{\frac{M^3-M}{6} - 2T_2}} \tag{2}$$

em que

$$S(d^2) = \sum_{1 \leq l \leq M} (U_l - V_l)^2,\ T_1 = \frac{\sum_{1 \leq i \leq k} (M_i^3 - M_i)}{12},\ T_2 = \frac{\sum_{1 \leq j \leq k} \left(N_j^3 - N_j\right)}{12}. \tag{3}$$

O coeficiente Tau de Kendall, T_{K}, para comparar preordenações, tem por base a noção de grafo de uma preordenação, $gr(\omega_i)$, em que

$$gr(\omega_i) = \left\{(\{x, y\}, \{z, t\}) \in F \times F : \{x, y\} \neq \{z, t\}, \{x, y\} \leq \{z, t\}\ e\ \{z, t\} > \{x, y\}\right\}$$

Sendo ω_1 e ω_2 preordenações totais tem-se que

$$|gr(\omega_1)| = \binom{M}{2} - \sum_{i=1}^{k} \binom{M_i}{2} = \sum_{i<j} M_i M_j$$

e

$$|gr(\omega_2)| = \binom{M}{2} - \sum_{i=1}^{h} \binom{N_i}{2} = \sum_{i<j} N_i N_j$$

Finalmente vem

$$T_k = \frac{C - D}{|gr(\omega_1)||gr(\omega_2)|} = \frac{C - D}{\sqrt{\sum_{i<j} M_i M_j}\sqrt{\sum_{i<j} N_i N_j}}. \tag{4}$$

O coeficiente de Goodman-Kruskal, T_{GK}, para comparar preordenações é dado por

$$T_{Gk} = \frac{C - D}{C + D}. \tag{5}$$

Este coeficiente é especialmente útil na comparação de preordenações que têm empates muito extensos, como os que ocorrem frequentemente nas preordenações associadas a hierarquias aquando da junção de duas classes numerosas. Outras

propriedades destes coeficientes e detalhes sobre as suas distribuições assintóticas podem ser vistas em Kendall (1970), Sousa (2000) ou Tendeiro (2005).

3.3 Exemplo

Considerem-se os dendrogramas (a) e (b) da Figura 6, onde $m = 5$ e $M = 10$. A preordenação associada ao dendrograma (a) é a preordem de F, do tipo (1,1,4,4), dada por $\{e_2,e_3\}<\{e_4,e_5\}<\{e_2,e_4\}=\{e_2,e_5\}=\{e_3,e_4\}=\{e_3,e_5\}<\{e_1,e_2\}=\{e_1,e_3\}=\{e_1,e_4\}=\{e_1,e_5\}$. Esta preordenação relaciona todos os pares de elementos por ordem crescente de semelhança entre si. Assim, por exemplo, $\{e_4,e_5\}<\{e_1,e_3\}$ significa que os elementos e_4 e e_5 são mais semelhantes entre si do que os elementos e_1 e e_3.

A preordenação associada ao dendrograma (b) é do tipo (1,1,2,6) e é dada por $\{e_2,e_3\}<\{e_4,e_5\}<\{e_1,e_2\}=\{e_1,e_3\}<\{e_1,e_4\}=\{e_1,e_5\}=\{e_2,e_4\}=\{e_2,e_5\}=\{e_3,e_4\}=\{e_3,e_5\}$.

(a) (b)

Figura 6 – Dendrogramas.

Escrevendo o conjunto F na forma $F=\{\{e_1,e_2\},\{e_1,e_3\},\{e_1,e_4\},\{e_1,e_5\},\{e_2,e_3\},\{e_2,e_4\},\{e_2,e_5\},\{e_3,e_4\},\{e_3,e_5\},\{e_4,e_5\}\}$, obtêm-se, para as variáveis U e V, os valores que constam da Tabela 1.

Tabela 1 – Valores das variáveis aleatórias U e V no exemplo considerado.

$\{e_i,e_j\}$	$\{e_1,e_2\}$	$\{e_1,e_3\}$	$\{e_1,e_4\}$	$\{e_1,e_5\}$	$\{e_2,e_3\}$	$\{e_2,e_4\}$	$\{e_2,e_5\}$	$\{e_3,e_4\}$	$\{e_3,e_4\}$	$\{e_4,e_5\}$
k	1	2	3	4	5	6	7	8	9	10
u_k	8.5	8.5	8.5	8.5	1	4.5	4.5	4.5	4.5	2
v_k	3.5	3.5	7.5	7.5	1	7.5	7.5	7.5	7.5	2

Aplicando as fórmulas (2) e (3) de cálculo de $S(d^2)$, T_1, T_2 e R_S é fácil verificar que $s(d^2) = 88$, $t_1 = 10$ e $t_2 = 18$, donde $r_S = 0.358$ (3 c.d.).

Atendendo a que $c = 20$, $d = 8$, $\sum_{i<j} M_i M_j = 33$ e $\sum_{i<j} N_i N_j = 29$, substituindo em (4) e (5), obtém-se $t_K = 0.388$ (3 c.d.) e $t_{GK} = 0.429$ (3 c.d.).

4 Distribuições dos coeficientes de comparação

Frequentemente é necessário interpretar probabilisticamente os valores fornecidos pelos coeficientes de comparação atrás definidos, por exemplo determinando a significância estatística. Alguns resultados sobre as distribuições assintóticas destes coeficientes existem (Kendall, 1970). O coeficiente de correlação ordinal de Spearman, R_S, tem distribuição assintótica normal de média zero e variância $\frac{1}{M-1}$. Em Kendall (1970) demonstra-se que a variável (C-D) tem também distribuição assintótica normal com:

$$E(C-D)=0 \quad \text{e} \quad V(C-D)=$$

$$=\frac{1}{18}\left[M(M-1)(2M+5)-\sum_{1\leq i\leq k} M_i(M_i-1)(2M_i+5)-\sum_{1\leq l\leq h} N_l(N_l-1)(2N_l+5)\right]$$

$$+\frac{1}{9M(M-1)(M-2)}\left[\sum_{1\leq i\leq k} M_i\,(M_i-1)(M_i-2)\right]\times\left[\sum_{1\leq l\leq h} N_l\,(N_l-1)(N_l-2)\right]$$

$$+\frac{1}{2M(M-1)}\left[\sum_{1\leq i\leq k} M_i(M_i-1)\right]\times\left[\sum_{1\leq l\leq h} N_l(N_l-1)\right] \qquad (6)$$

Para o coeficiente de Goodman-Kruskal há também alguns resultados assintóticos para a distribuição normal, para situações particulares, que pela sua complexidade e não utilização no presente trabalho não serão aqui apresentados.

O recurso às distribuições assintóticas destes coeficientes está contudo comprometido no contexto do presente trabalho, principalmente por dois motivos:

1. As estruturas de classificação a comparar são obtidas a partir de um mesmo quadro de dados.
2. A propriedade ultramétrica induz relações entre os valores das sequências a comparar.

Assim a hipótese de "independência", subjacente à dedução das distribuições assintóticas, não se verifica quando se trata da comparação de estruturas de classificação.

O objectivo natural seria a obtenção das correspondentes distribuições exactas para estes coeficientes, aplicados à comparação de estruturas de classificação. Foi já referido que, na aplicação de uma C.H.A. a um conjunto de dados, a informação contida no dendrograma, na matriz ultramétrica ou na hierarquia indiciada, é equivalente, e as estruturas de classificação associadas a estas entidades são coincidentes. Vamos, por isso, centrar-nos em dendrogramas. A obtenção da distribuição exacta de um coeficiente neste contexto passaria pela enumeração do

conjunto de todos os dendrogramas, fixado o número de nós terminais. Contudo tal não é exequível, tendo em conta o rápido crescimento de d(*m*) (ver (1)):

Tabela 2 – Número de dendrogramas não isomorfos.

m	4	5	6	10	15	20	50
d(*m*)	18	180	2700	2571912000	5.14×10^{18}	1.53×10^{29}	$>10^{100}$

A solução passa pela determinação de distribuições empíricas, por recurso à simulação.

Para a obtenção de amostras aleatórias de dendrogramas recorreu-se a três métodos de geração aleatória de dendrogramas propostos na literatura:

- método uniforme (Sousa, 2000)
- método RA (Podani, 2000)
- método da Permutação Dupla (Lapointe & Legendre, 1991).

Estes três métodos são uniformes no sentido de Furnas (1984), isto é, para um valor de *m* fixo, os métodos geram todos os dendrogramas com *m* nós terminais de forma equiprovável (com probabilidade 1/d(*m*)). Os valores de *m* contemplados no estudo foram: 4(1)15, 20(5)50, 75, 100(100)500. A metodologia seguida pode resumir-se em três passos:

1. gerar um par de dendrogramas
2. calcular o valor do coeficiente de comparação
3. repetir 1. e 2. *k* vezes.

No nosso caso fez-se $k = 1000$.

Sendo os três métodos de geração aleatória de dendrogramas todos uniformes no sentido de Furnas, os resultados obtidos por esta metodologia devem ser semelhantes, independentemente do método de geração aleatória que se use no passo 1.. Contudo optou-se por utilizá-los todos, meramente a título comprovativo.

Os algoritmos foram implementados em linguagem Fortran. O gerador de números pseudoaleatórios baseia-se na subrotina "random_number"; a semente dos números aleatórios foi sempre gerada automaticamente pelo processador de acordo com o relógio do mesmo.

5 Discussão dos resultados

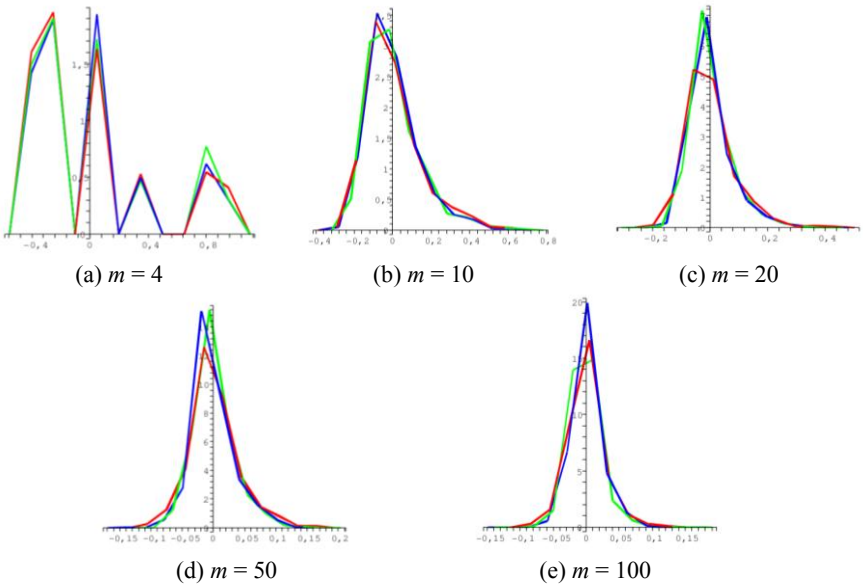
A apresentação e discussão dos resultados são, por razões de limitação de número de páginas, feitas apenas para o coeficiente Tau de Kendall. Refira-se que os

resultados encontrados para os outros coeficientes (podem ser consultados em Tendeiro, 2005) são análogos. A Figura 7 (a)-(e) apresenta, para diferentes valores de m, os polígonos de frequências dos valores do coeficiente Tau de Kendall correspondentes aos três métodos de geração. De facto, verifica-se que os polígonos são bastante semelhantes entre si, como seria de esperar.

A escolha do método de geração aleatória de dendrogramas é irrelevante. Assim, sem prejuízo das conclusões que nos propomos tirar, vamos usar a distribuição empírica obtida através do método uniforme.

Uma primeira inspecção aos resultados obtidos permite retirar algumas conclusões:

- os valores da amplitude amostral, bem como da dispersão quartal, são cada vez menores conforme m aumenta; este facto está relacionado com o rápido crescimento de d(m);
- para cada valor de m verifica-se que a mediana é quase sempre negativa, aproximando-se de zero conforme m aumenta;
- as distribuições são assimétricas positivas (os gráficos apoiam este facto), sendo que a assimetria vai diminuindo conforme m aumenta.



(a) $m = 4$ (b) $m = 10$ (c) $m = 20$

(d) $m = 50$ (e) $m = 100$

Figura 7 – Polígonos de frequências para o coeficiente Tau de Kendall.

A comparação entre as distribuições empíricas obtidas e a distribuição assimptótica é pertinente. No resultado da distribuição assintótica do coeficiente de Kendall, Secção 4, são dadas as expressões para as média e variância de C-D.

Sendo C-D uma variável cuja variação de valores depende do comprimento da preordenação e dos seus patamares de empates, a interpretação dos seus valores é mais delicada e fica inviabilizada uma comparação gráfica. No que respeita à média tem-se que E(C-D)=0 e, consequentemente, $E(T_K)=0$. Na Tabela 3 apresentam-se, para os valores de *m* considerados, os valores empíricos das médias de C-D e de T_K. No que respeita à variância apresentam-se, para a variável C-D, os valores para a distribuição assintótica (V_t), para a distribuição empírica (V_a), bem como uma comparação relativa desses valores.

Tabela 3 – Análise comparativa da média e dispersão das distribuições empírica e assintótica.

m	Média amostral		Variância de C-D		
	C-D	T_K	Amostral (V_a)	Teórica V_t	$\frac{V_a - V_t}{V_t}$
4	.2400	$.2427\text{x}10^{-1}$	$.1870\text{x}10^{2}$	$.1797\text{x}10^{2}$	.0406
5	-.5500	$-.1702\text{x}10^{-2}$	$.6881\text{x}10^{2}$	$.8674\text{x}10^{2}$	-.2068
6	-.1100	$-.6648\text{x}10^{-3}$	$.3382\text{x}10^{3}$	$.2960\text{x}10^{3}$	.1426
7	-.1190x10	$-.5755\text{x}10^{-2}$	$.9325\text{x}10^{3}$	$.8284\text{x}10^{3}$	.1256
8	-.5520x10	$-.2052\text{x}10^{-1}$	$.2478\text{x}10^{4}$	$.1989\text{x}10^{4}$	.2459
9	$-.1631\text{x}10^{2}$	$-.3397\text{x}10^{-1}$	$.3960\text{x}10^{4}$	$.4240\text{x}10^{4}$	-.0658
10	$.1702\text{x}10^{2}$	$.2278\text{x}10^{-1}$	$.1217\text{x}10^{4}$	$.8252\text{x}10^{4}$	.4749
11	-.9140x10	$-.8627\text{x}10^{-2}$	$1902\text{x}10^{5}$	$.1494\text{x}10^{5}$	.2728
12	.2600x10	$.1702\text{x}10^{-2}$	$.2728\text{x}10^{5}$	$.2590\text{x}10^{5}$	.0533
13	$.2225\text{x}10^{2}$	$.1028\text{x}10^{-1}$	$.5474\text{x}10^{5}$	$.4293\text{x}10^{5}$	.2753
14	$.5771\text{x}10^{2}$	$.1835\text{x}10^{-1}$	$.1193\text{x}10^{6}$	$.6886\text{x}10^{5}$	.7330
15	$-.6442\text{x}10^{2}$	$-.1550\text{x}10^{-1}$	$.1342\text{x}10^{6}$	$.1058\text{x}10^{6}$	.2696
20	$-.1033\text{x}10^{3}$	$-.7201\text{x}10^{-2}$	$.1002\text{x}10^{7}$	$.6273\text{x}10^{6}$	.5976
25	$.1993\text{x}10^{3}$	$.5800\text{x}10^{-2}$	$.4339\text{x}10^{7}$	$.2497\text{x}10^{7}$	.7379
30	$-6190\text{x}10^{2}$	$-.9200\text{x}10^{-3}$	$.1317\text{x}10^{8}$	$.7593\text{x}10^{7}$	.7350
35	$-.8816\text{x}10^{3}$	$-.6083\text{x}10^{-2}$	$.4605\text{x}10^{8}$	$.1975\text{x}10^{8}$	.1332x10
40	$-.1565\text{x}10^{3}$	$-.5496\text{x}10^{-3}$	$.1385\text{x}10^{9}$	$.4421\text{x}10^{8}$	.2132x10
45	$-.3524\text{x}10^{4}$	$-.9068\text{x}10^{-2}$	$.2972\text{x}10^{9}$	$.9142\text{x}10^{8}$	.2251x10
50	$-.1349\text{x}10^{3}$	$-.2047\text{x}10^{-3}$	$.7178\text{x}10^{9}$	$.1709\text{x}10^{9}$	.3201x10
75	$.2075\text{x}10^{5}$	$.6839\text{x}10^{-2}$	$.1152\text{x}10^{11}$	$.1990\text{x}10^{10}$	.4789x10
100	$.1010\text{x}10^{5}$	$.9444\text{x}10^{-3}$	$.7153\text{x}10^{11}$	$.1142\text{x}10^{11}$	.5263x10
200	$.2968\text{x}10^{6}$	$.1846\text{x}10^{-2}$	$.5946\text{x}10^{13}$	$.7405\text{x}10^{12}$	.7030x10
300	$.1897\text{x}10^{7}$	$.2406\text{x}10^{-2}$	$.9577\text{x}10^{14}$	$.8475\text{x}10^{13}$	$.1030\text{x}10^{2}$
400	$-.3218\text{x}10^{6}$	$-.8538\text{x}10^{-3}$	$.8677\text{x}10^{15}$	$.4800\text{x}10^{14}$	$.1708\text{x}10^{2}$
500	$.4264\text{x}10^{7}$	$.6535\text{x}10^{-3}$	$.4229\text{x}10^{16}$	$.1803\text{x}10^{15}$	$.2246\text{x}10^{2}$

De notar que o valor da variância teórica, dado por (6), é função do comprimento dos patamares da preordenação, não sendo, por isso, constante para cada valor de m. Os valores da variância teórica (coluna V_t na Tabela 3) foram calculados de acordo com o seguinte: (i) para cada par de dendrogramas gerado determinou-se o comprimento dos respectivos patamares e o correspondente valor de variância teórica (fórmula (6)) (ii) para cada valor de m calculou-se a média dos valores de variância teórica obtidos.

Os valores da média amostral de T_k oscilam em torno de zero, valor da média teórica, registando-se uma tendência para um decréscimo dos valores em módulo à medida que m cresce. No que respeita à dispersão verifica-se que os valores empíricos são quase sempre superiores aos valores teóricos. Tratando-se da variável C-D, os valores da variância, amostrais e teóricos, são muito elevados. Para uma análise comparativa consideram-se os valores da diferença relativa das variâncias. Constata-se que a diferença relativa aumenta com m, assumindo valores inferiores a 1 quando $m \leq 30$ e passando a valores superiores a 1 quando $m \geq 35$. Estes factos parecem permitir concluir que a utilização da distribuição assintótica, no contexto da comparação de estruturas de classificação, é desadequada para qualquer valor de m.

6 Conclusão

Neste trabalho é apresentada uma metodologia para comparar quantitativamente pares de estruturas de classificação, que passa por associar a cada estrutura de classificação uma preordenação sobre o conjunto de elementos a classificar. São apresentados coeficientes de correlação (associação) ordinal para a comparação de pares de preordenações e discutida a adequação das suas distribuições teóricas ao problema em estudo, já que os pressupostos de independência, habitualmente aceites, não são aqui verificados.

Propõe-se a dedução de distribuições empíricas obtidas por simulação. A metodologia proposta é ilustrada com o caso em que as estruturas a comparar são dendrogramas, obtidos a partir do mesmo conjunto de dados. Para a dedução das distribuições empíricas recorreu-se a métodos de geração aleatória de dendrogramas. Os resultados obtidos parecem confirmar que os problemas teóricos observados na Secção 4 limitam, de facto, a utilização das distribuições assintóticas dos coeficientes de comparação no contexto da comparação de dendrogarmas. São apresentados e discutidos resultados para o coeficiente Tau de Kendall.

Várias aplicações, como intervalos de confiança ou testes de hipóteses, podem ser levadas a cabo a partir do conhecimento das distribuições empíricas dos coeficientes aqui estudados.

Referências

BOCK, H. H. (1996). Probability Models and Hypotheses Testing in Partitioning Cluster Analysis, in *Clustering and Classification* (eds. P. Arabie, L. J. Hubert, G. De Soete), World Scientific, Singapore, 377-453.

FURNAS, G. W. (1984). The Generation of Random, Binary Unordered Trees, *Journal of Classification*, 1, 187-233.

GORDON, A. D. (1996). Hierarchical Classification, in *Clustering and Classification* (eds. P. Arabie, L. J. Hubert, G. De Soete), World Scientific, Singapore, 65-121.

GORDON, A. D. (1999). *Classification*, 2nd Edition, Chapman & Hall, London.

HALKIDI, M.; SBATISTAKIS, Y. e VAZIRGIANNIS, M. (2001). On Clustering Validation Techniques, *Journal of Intelligent Information Systems*, 17: 2/3, 107-145.

KENDALL, M. G. (1970). *Rank Correlation Methods*, 4ª Edição, Griffin, London.

LAPOINTE, F. J. e LEGENDRE, P. (1991). The Generation of Random Ultrametric Matrices Representing Dendrograms, *Journal of Classification*, 8, 177-200.

NICOLAU, F. C. (1984). Problemas de Validade em Classificação Automática, *Actas do III Colóquio de Estatística e Investigação Operacional*, SPEIO, Lagos.

PODANI, J. (2000). Simulation of Random Dendrograms and Comparison Tests: Some Comments, *Journal of Classification*, 17, 123-142.

SOUSA, F. (2000). *Novas Metodologias em Classificação Hierárquica Ascendente*, Dissertação de Doutoramento, Universidade Nova de Lisboa, Lisboa.

SOUSA, F. e NICOLAU, F. C. (2001). Uma Abordagem ao Problema da Comparação de Estruturas Classificatórias, em *A Estatística em Movimento* (M. M. Neves, J. Cadima, M. J. Martins, F. Rosado), Sociedade Portuguesa de Estatística, 409-418.

TENDEIRO, J. (2005). *Comparação de Dendrogramas: Obtenção de Distribuições Empíricas de Alguns Coeficientes*, Dissertação de Mestrado, Universidade do Porto, Porto.