Avaliação de Algumas Medidas de Dissimilaridade de Simetria Ajustável

Alberto Sampaio[1]*, Isabel B. Sampaio [2], Gustavo Alves[3]

[1]Departamento de Engenharia Informática e LABORIS (Research Laboratory on Systems), Instituto Superior de Engenharia do Porto, Instituto Politécnico do Porto, Portugal, as@dei.isep.ipp.pt

[2]Departamento de Engenharia Informática, Instituto Superior de Engenharia do Porto, Instituto Politécnico do Porto, Portugal, is@dei.isep.ipp.pt

[3]Departamento de Engenharia Electrotécnica e LABORIS (Research Laboratory on Systems), Instituto Superior de Engenharia do Porto, Instituto Politécnico do Porto, Portugal, gca@isep.ipp.pt

Resumo

As medidas, ou coeficientes, de similaridade utilizados habitualmente no cálculo da proximidade entre duas entidades destinam-se a dados simétricos, ou a dados assimétricos. Neste artigo são apresentados coeficientes de similaridade de simetria ajustável, propostos recentemente, e é avaliada a distorção que produzem numa classificação, usando vários critérios de aglomeração hierárquica aplicados e três conjuntos diferentes de dados. Estendeu-se a avaliação feita anteriormente a três conjuntos de dados. Na avaliação compararam-se os valores obtidos com uma versão simplificada desses coeficientes com os obtidos usando os correspondentes coeficientes simétricos e assimétricos. Os resultados obtidos com os coeficientes de simetria ajustável foram superiores aos obtidos com os outros dois tipos de coeficientes.

Palavras Chave: medidas de (di)similaridade, simetria, simetria ajustável, classificação.

1. Introdução e Antecedentes

Podemos distinguir entre dois tipos de variáveis (caracteres) dicotómicas, as *simétricas* e as *assimétricas* (também chamadas *não-simétricas*). Nas variáveis simétricas é indiferente qual o estado que deve ser codificado com 0 ou 1. É o caso do atributo sexo, cujos estados "Homem" e "Mulher" podem ser codificados indistintamente com 0 ou 1. É por esta razão que as variáveis simétricas também são conhecidas como *invariantes*. O mesmo não acontece, por exemplo, com a variável "É atleta de alta competição" que, num estudo hipotético sobre o efeito do desporto na saúde, permitiria saber que duas pessoas com esse atributo teriam muito em comum, mas pouco ou nada diria no caso da ausência desse atributo. Ou seja, trata-se de uma variável assimétrica. Nestes casos, os pares discordantes são normalmente ignorados no cálculo da (dis)similaridade entre duas entidades. Como se percebe, a diferença entre os coeficientes simétricos e assimétricos reside, respectivamente, na utilização, ou não, do número de pares nulos, *d*. Têm sido propostas medidas (ou coeficientes) de similaridade para ambos os tipos de variáveis.

Recentemente, em [Sampaio e Sampaio, 2006], propusemos uma solução alternativa, que consiste em encontrar um parâmetro que ajuste a medida de similaridade à quantidade de assimetria que se saiba, ou estime, existir nos dados. A solução proposta associa o coeficiente de simetria (*CS*) abaixo em (1), a *d*, em que *n* é o número de caracteres e *NS* (Nível de asSimetria ou Simetria) um parâmetro que variará proporcionalmente à quantidade de assimetria global presente nos dados.

$$CS = \frac{n - NS}{n} \qquad (1)$$

O parâmetro *NS* pode tomar valores no intervalo [0 , *n* [. Teoricamente *NS* nunca poderá ser *n*. O termo *d* só se anula se o coeficiente for nulo, o que acontece quando *NS* é igual a *n*, e nesse caso obter-se-ia um coeficiente assimétrico. Se todos os dados forem simétricos, *NS* terá valor 0 , obtendo-se uma medida de similaridade simétrica. De uma forma simplificada *NS* também poderia ser visto como variando com o número de caracteres assimétricos.

Aplicando o coeficiente *CS* a medidas de (dis)similaridade já existentes, resultará um conjunto de coeficientes (medidas) a que chamámos medidas de simetria ajustável. Na Tabela 1 são mostrados a título de exemplo três dessas medidas de dissimilaridade com indicação dos coeficientes de similaridade a partir dos quais foram obtidos, subtraindo-se a 1 para transformação em dissimilaridade.

Tabela 1 Algumas medidas de dissimilaridade de simetria ajustável.

Coeficiente de Simetria Ajustável		Baseado em
$\frac{b+c}{a+b+c+d\frac{n-NS}{n}}$	(asm)	"Simple Matching Coefficient"/ Jaccard (ssm)
$\frac{2*(b+c)}{a+d\frac{n-NS}{n}+2*(b+c)}$	(art)	Roger e Tanimoto/Dice (srt)
$\frac{b+c}{2*\left(a+d\frac{n-NS}{n}\right)+b+c}$	(ass)	Sokal e Sneath (sss)

No caso em que não se saiba qual o número de caracteres assimétricos, e/ou a sua simetria, parece razoável considerar metade dos dados simétricos, com um nível intermédio também de simetria, e os coeficientes seriam considerados semi-simétricos.

No caso mais geral, como o nível de assimetria também pode variar de carácter para carácter, então, nesse caso, *NS* deverá ser igual à soma dos pesos das assimetrias do subconjunto dos caracteres assimétricos, com o peso da assimetria de cada carácter a variar entre 0 e 1 consoante a assimetria do carácter.

2. Método de Avaliação

Com a avaliação pretendeu-se estudar a capacidade de as medidas de simetria ajustável produzirem classificações de qualidade superior às suas correspondentes simétricas e assimétricas. Para isso, procedeu-se à comparação de medidas dos três tipos considerados: simétricas, assimétricas e de simetria ajustável. O processo de avaliação utilizado foi o seguinte: a) aplicação de várias medidas dos diferentes tipos aos conjuntos de dados; b) aplicação de quatro critérios de aglomeração hierárquica: Ligação Mínima (*Single Linkage*), Ligação Máxima (*Complete Linkage*), Ligação Média (*Average*) e Método *Ward* a cada uma das matrizes de dissimilitude obtidas no passo anterior para cada uma das medidas; c) medição da qualidade de cada uma das classificações obtidas; d) soma dos valores anteriores e sua comparação entre os três tipos de medidas. Houve ainda necessidade de ajustar as medidas de simetria *CS* aos dados de forma a melhorar o desempenho das respectivas medidas.

Coeficientes avaliados

Os coeficientes (medidas) de dissimilaridade utilizados foram: os simétricos e assimétricos definidos na Tabela 2; e os correspondentes de simetria ajustável. Estes últimos foram utilizados inicialmente na sua versão simplificada com um valor de *CS*=1/2, logo, *d*/2. Todos estas medidas foram obtidas após transformação em dissimilaridades das correspondentes similaridades, usando $\sqrt{1-S}$ para a transformação, sendo *S* o coeficiente de similaridade.

Tabela 2 Medidas de dissimilaridade simétricas e assimétricas.

Simétricos	Assimétricos
"Simple Matching Coefficient" (1938) (ssm)	Coeficiente de Jacard (1908) (nja)
$\frac{b+c}{a+b+c+d}$	$\frac{b+c}{a+b+c}$
Roger e Tanimoto (1960) (srt)	Dice (1945), Sorensen (1948) (nds)
$\frac{2*(b+c)}{a+d+2*(b+c)}$	$\frac{b+c}{2*a+b+c}$
Sokal e Sneath (1963) (sss)	Sokal e Sneath (1963) (nss)
$\frac{b+c}{2*(a+d)+b+c}$	$\frac{2*(b+c)}{a+2*(b+c)}$

Estatística

A qualidade das classificações resultantes foi medida através do valor do coeficiente de correlação cofenética (CCC) [Sneath e Sokal, 1973] obtido com cada par coeficiente de dissimilaridade/critério de aglomeração.

Dados utilizados

Pretendeu-se incluir diferentes conjuntos de dados, em que pelo menos um deveria conter apenas variáveis consideradas assimétricas. Um conjunto (A) de dados disponível usado em [Sampaio et al., 2005] continha 112 variáveis de tipos diferentes, pelo que houve que necessidade de as transformar todas em dicotómicas (usou-se o valor médio para esta transformação). Outro conjunto (B), consiste num subconjunto das primeiras cinco variáveis do anterior. Um conjunto (C) obtido da Tabela 12 (p. 33) de [Kaufman e Rousseeuw, 1990], retendo apenas os dados das três variáveis assimétricas.

3. Estudo e Resultados

Os coeficientes de simetria ajustável foram utilizados inicialmente com um *CS*=1/2, mas a soma dos valores da correlação cofenética para esses coeficientes era superior à obtida com os simétricos, mas inferior à obtida com os assimétricos. Por isso, procurou analisar-se o comportamento dos coeficientes ajustáveis para outros níveis de simetria, fazendo variar o denominador de *d*, num processo que foi exploratório.

Na Tabela 3 encontram-se os resultados obtidos para a soma dos doze valores dos coeficientes de correlação cofenética, para cada um dos três conjuntos de dados analisados e, no caso dos coeficientes ajustáveis, para o valor mais favorável de CS.

Tabela 3 Resumo das somas dos valores CCC.

Conjunto	Simétrico	(*CS*) Ajustável		Assimétrico
A	10,14638	(1/80)	10,97312	10,79017
B	11,32328	(1/4)	11,41974	11,34903
C	10,56276	(1/7)	11,02306	11,01015

Nesta tabela, dentro de parênteses encontram-se os valores utilizado para o coeficiente *CS*.

4. Discussão e Conclusões

Os coeficientes de simetria ajustável obtiveram globalmente valores mais elevados para o coeficiente de correlação cofenética para cada um dos conjuntos de dados. Daqui, pode-se concluir que os coeficientes de simetria ajustável produziram classificações mais fieis às matrizes de proximidade do que os seus análogos simétricos e assimétricos. Neste estudo houve o cuidado de incluir um conjunto de dados assimétricos e verificou-se que também para esse conjunto de dados assimétrico o resultado foi favorável

aos coeficientes ajustáveis. Na utilização dos coeficientes de simetria ajustável não houve preocupação em procurar o valor óptimo do denominador, porque se tratava apenas de saber se esses coeficientes permitiam em geral melhores classificações que os outros coeficientes.

Também para os três conjuntos, os coeficientes assimétricos foram superiores aos simétricos, o que está de acordo com a ideia expressa por Anderberg [1973] de que os pares nulos não acrescentariam qualquer informação útil para uma classificação. Este estudo ajuda a confirmar essa ideia quando se consideram apenas os coeficientes simétricos e assimétricos, mas, também de acordo com os resultados obtidos neste estudo, uma presença parcial dos pares nulos parece ser preferível à sua completa ausência. Também se observou que o valor de *CS* para um valor óptimo da soma dos CCC varia com o conjunto de dados utilizado. Também é de notar que, contrariamente ao que seria de esperar, o conjunto *C* não foi aquele em que se observou uma diferença mais acentuada (A-0,64362; B-0,02575; C-0,447382) entre as somas dos CCC's dos coeficientes simétricos e assimétricos. É possível que tal se possa explicar pelas dimensões diferentes dos conjuntos de dados.

Algumas limitações do estudo foram a ausência de validação estatística dos resultados, a junção dos coeficientes com os critérios, o que limita as conclusões, a utilização de apenas um índice de qualidade da classificação, baseado na correlação produto-momento, e, a análise realizada individualmente a cada coeficiente estar por fazer.

Como trabalho futuro imediato pretende-se analisar individualmente cada coeficiente. Também se prevê alargar o número de coeficientes a comparar, utilizar outras estatísticas para avaliação da qualidade das soluções obtidas com os vários coeficientes para avaliação do impacto da variação do nível de assimetria nos resultados das próprias estatísticas e vice-versa. Dever-se-á também analisar a relação entre os dados de um estudo e o nível de assimetria a indicar nos coeficientes. Igualmente importante será estudar a adequação destes coeficientes na estimação do nível de assimetria dos dados. Finalmente, propomo-nos fazer a análise teórica destes coeficientes e suas relações com outros trabalhos da mesma índole.

Referências

[1] Anderberg, M.R., Cluster Analysis for Applications, Academic Press, 1973.

[2] Gower, J.C., A general Coefficient of Similarity and Some of Its Properties, Biometrics, Vol.27, pp.857-871, 1971.

[3] Kaufman, L., Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, Inc., 1990.

[4] Sampaio, A., Gray, E.M., Martins, M., A Comparison of SPA Methods, 31ts EUROMICRO Conference on Software Engineering and Advanced Applications (SEAA), WiP session , August 31st-September 3rd, 2005, Porto, Portugal, 2005.

[5] Sampaio, A., Sampaio, I.B., Proposta de Coeficientes de Associação de Simetria Ajustável e sua Avaliação Parcial, XIV Congresso Anual da Sociedade Portuguesa de Estatística, 27 a 30 de Setembro de 2006, Covilhã, Portugal, 2006.

[6] Sneath, P.H.A., Sokal, R.R., Numerical Taxonomy: The Principles and Practice of Numerical Classification, W.H. Freeman and Company, San Francisco, USA, 1973.