JOCLAD XXI

Jornadas de Classificação e Análise de Dados 2014

LIVRO DE RESUMOS

10 - 12 Abril | Lisboa - Portugal

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

CLAD
Associação Portuguesa de Classificação e Análise de Dados
20 Anos

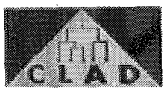
Reunião Anual da

ASSOCIAÇÃO PORTUGUESA DE CLASSIFICAÇÃO

E ANÁLISE DE DADOS (CLAD)

Livro de Resumos

AS JOCLAD 2014 TIVERAM O APOIO INSTITUCIONAL DE:



Associação Portuguesa de Classificação e Análise de Dados

INSTITUTO NACIONAL DE ESTATÍSTICA

STATISTICS PORTUGAL

Ficha Técnica

Presidente das Jornadas

Alda Carvalho (Presidente do INE)

Secretário das Jornadas

Fernanda Sousa (Presidente da CLAD e FEUP-Universidade do Porto)

Comissão Organizadora

Catarina Marques (ISCTE-Instituto Universitário de Lisboa)

Isabel Silva (FEUP-Universidade do Porto)

José Gonçalves Dias (ISCTE-Instituto Universitário de Lisboa)

Nuno Lavado (ISEC-Instituto Politécnico de Coimbra)

Título: XXI Jornadas de Classificação e Análise de Dados (JOCLAD 2014). Livro de Resumos.

Produzido: Instituto Nacional de Estatística

Editores: Fernanda Sousa, Catarina Marques, Isabel Silva, José Gonçalves Dias, Nuno Lavado, Carlos Marcelo

ISBN: **978-989-98955**

ST III – Sessão 20 anos da CLAD – 6ª Feira, 11 de Abril, Salão Nobre (17h05)

Classes de objectos simbólicos: dados da indústria automóvel

Áurea Sousa [1], Helena Bacelar-Nicolau [2], Fernando C. Nicolau [3], Osvaldo Silva [4]
Universidade dos Açores, Departamento de Matemática, CEEAplA, CMATI, aurea@uac.pt;
Universidade de Lisboa, Faculdade de Psicologia (LEAD), ISAMB, hbacelar@fp.ul.pt;
Univ. Nova de Lisboa, FCT, Dep. de Matemática e DataScience, geral@datascience.org
Universidade dos Açores, Departamento de Matemática, CES, CMATI, osilva@uac.pt

Sumário

Neste trabalho, é abordada a Análise Classificatória Hierárquica Ascendente (ACHA) de dados simbólicos ou complexos (generalizações de dados clássicos), com base no coeficiente de afinidade generalizado ponderado e em critérios de agregação clássicos e probabilísticos, estes últimos no âmbito da metodologia *VL*. São apresentados os principais resultados obtidos com a ACHA de 33 modelos de carros (dados simbólicos na área da indústria automóvel), com base no coeficiente de afinidade generalizado ponderado, centrado e reduzido pelo método de Wald e Wolfowitz, comparando-se os resultados obtidos com os de outros autores e com a partição definida *a priori* pelas categorias (*"Utilitário", "Berlina", "Desportivo", "Luxo")* a que os modelos de carros pertencem.

Palavras-chave: Análise classificatória hierárquica, Coeficiente de afinidade generalizado ponderado, Dados simbólicos ou complexos, Metodologia *VL*.

1. Introdução

Na sociedade actual, onde os avanços computacionais têm imperado, é cada vez mais frequente a utilização de bases de dados, sendo fundamental efectuar a síntese de conjuntos de dados de elevada dimensão em termos dos seus conceitos subjacentes, os quais têm de ser descritos por dados mais complexos, designados por dados simbólicos. Estes dados podem ser heterogéneos e são representados em tabelas, cujas células podem conter um ou mais valores, tais como subconjuntos de categorias, intervalos da recta real, ou distribuições de frequências (Bock and Diday, 2000; Bacelar-Nicolau, 2000, 2002; Bacelar-Nicolau et al., 2009, 2010; Sousa et al., 2013). As linhas da tabela de dados representam unidades de dados ou objectos simbólicos e as colunas representam variáveis simbólicas.

Muitas medidas de proximidade entre objectos simbólicos têm sido referidas na literatura (Bock e Diday, 2000). Uma vez obtida a matriz de proximidades entre os elementos do conjunto a classificar, podem ser aplicados critérios de agregação clássicos ou probabilísticos (Bacelar-Nicolau et al., 2009, 2010; Sousa et al., 2013). Neste trabalho, a ACHA foi efectuada com base no coeficiente de afinidade generalizado ponderado (Bacelar-Nicolau, 2000; Bacelar-Nicolau et al., 2009, 2010) e em três critérios de agregação probabilísticos no âmbito da Metodologia *VL* (Nicolau, 1983; Nicolau e Bacelar-Nicolau, 1998), sobre um conjunto de dados retirado da literatura da análise de dados complexos.

2. Análise classificatória hierárquica de objectos simbólicos com base no coeficiente de afinidade generalizado ponderado

A Análise Classificatória (*Cluster Analysis*) tem como objectivo identificar grupos (classes) de entidades (indivíduos, objectos, etc.), relativamente homogéneos e, de preferência, bem separados, com base nas semelhanças ou dissemelhanças entre essas entidades. Os métodos hierárquicos aglomerativos começam por considerar um número de classes igual ao número de elementos a classificar e posteriormente, em cada etapa, efectuam a junção ou aglomeração de classes em classes maiores, obtendo-se na última etapa um único grupo contendo todos os elementos a classificar.

A partir do coeficiente de afinidade entre duas distribuições de probabilidade discretas, proposto por Matusita (1951), Bacelar-Nicolau (1980, 1988) introduziu o coeficiente de afinidade no domínio da Análise Classificatória, para avaliar a semelhança básica entre pares de colunas ou pares de linhas de uma matriz de dados, ou seja, entre variáveis ou entre indivíduos, conforme o conjunto que se pretende classificar. Este coeficiente foi estendido a diferentes tipos de dados, incluindo dados de tipo heterogéneo e de natureza complexa (ou simbólicos) (Bacelar-Nicolau, 2000, 2002; Bacelar-Nicolau et al., 2009, 2010), frequentemente presentes em bases de dados de elevada dimensão. Os critérios de agregação probabilísticos usados, no âmbito da Metodologia *VL*, recorrem essencialmente a noções probabilísticas para a definição das funções de comparação (Nicolau, 1983; Nicolau e Bacelar-Nicolau, 1998). A extensão do coeficiente de afinidade para o caso de dados simbólicos é designada por coeficiente de afinidade generalizado ponderado (Bacelar-Nicolau, 2000, 2002; Bacelar-Nicolau et al., 2009, 2010). O coeficiente assintoticamente centrado e reduzido, sob uma hipótese de referência permutacional baseada no teorema limite de Wald e Wolfowitz permite, por sua vez, definir um coeficiente probabilístico no contexto da metodologia *VL*, na linha iniciada por Lerman (1972, 1981) e desenvolvida por Bacelar-Nicolau (e.g.1980, 1987, 1988) e Nicolau (e.g. 1983, 1998). Aplicações da extensão do coeficiente de afinidade para o caso de dados simbólicos foram apresentadas, por exemplo, em Bacelar-Nicolau et al. (2009, 2010) e em Sousa et al. (2013).

3. Exemplo da indústria automóvel: "*Car data set*"

A matriz de dados simbólicos que aqui usamos, para exemplificar a metodologia, é referida na literatura da análise de dados simbólicos (e.g., De Carvalho et al., 2006*a*, 2006*b*; Souza et al., 2007) e contém trinta e três modelos de carros (objectos simbólicos), descritos por oito variáveis cujos valores são intervalos da recta real ("*Preço*", "*Cilindrada*", "*Velocidade Máxima*", "*Aceleração*", "*Passo*", "*Comprimento*", "*Largura*" e "*Altura*"), duas variáveis categóricas ("*Alimentação*" e "*Tracção*") com categorias não ordenadas que podem assumir múltiplos valores (subconjuntos de categorias) e uma variável nominal ("*Categoria do Carro*"). Esta última variável, com as modalidades "*Utilitário*", "*Berlina*", "*Desportivo*" e "*Luxo*", reflecte a classificação *a priori* dos modelos de carros, a qual pode ser encontrada, por exemplo, em De Carvalho et al. (2006*a*, 2006*b*). A Tabela 1 mostra uma parte da matriz de

dados simbólicos, sendo de referir que a matriz de dados completa está disponível no software SODAS (Symbolic Official Data Analysis System).

Tabela 1-Parte da matriz de dados simbólicos -"Car data set"

Modelo	*Preço*	*Cilindrada*	*Alimentação*	*Tracção*	..	*Altura*	*Categoria*
Alfa 145	[27806, 33596]	[1370, 1910]	*Gasoli, Diese*	*Anter*	..	[143, 143]	*Utilit*
Alfa 156	[41593, 62291]	[1598, 2492]	*Gasoli*	*Anter*	..	[142, 142]	*Berlina*
...	...	...	...		..	...	...
Passat	[39676, 63455]	[1595, 2496]	*Gasoli, Diese*	*Anter, Integ*	..	[146, 146]	*Luxo*

Nesta comunicação, são apresentados os principais resultados obtidos com a *A.C.H.A.* dos trinta e três modelos de carros, com base no coeficiente de afinidade generalizado ponderado, centrado e reduzido pelo método de Wald e Wolfowitz, e em três critérios de agregação probabilísticos, *AVL*, *AVI* e *AVB* (Nicolau, 1983; Bacelar-Nicolau, 1988; Nicolau e Bacelar-Nicolau, 1998; Lerman, 1972, 1981). Os principais resultados obtidos são comparados com os de outros autores (e.g., De Carvalho *et al.*, 2006*a*, 2006*b*; Souza *et al.*, 2007).

Referências

BACELAR-NICOLAU, H. (1980) *Contribuições ao Estudo dos Coeficientes de Comparação em Análise Classificatória,* Tese de Doutoramento, FCL, Universidade de Lisboa.

BACELAR-NICOLAU, H. (1987) On the Distribution Equivalence in Cluster Analysis. In Devijver, P.A. & Kittler, J. (Ed.) Pattern Recognition Theory and Applications, NATO ASI Series, Series F: Computer and Systems Sciences, vol. 30, New York, Springer - Verlag, 73-79.

BACELAR-NICOLAU, H. (1988) Two Probabilistic Models for Classification of Variables in Frequency Tables. IN BOCK, H.-H. (Ed.) *Classification and Related Methods of Data Analysis.* North Holland, Elsevier Sciences Publishers B.V., pp. 181-186.

BACELAR-NICOLAU, H. (2000) The Affinity Coefficient. IN BOCK, H.-H. & DIDAY, E. (Ed.) *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data.* Series: Studies in Classification, Data Analysis, and Knowledge Organization, Berlin, Springer-Verlag, 160-165.

BACELAR-NICOLAU, H. (2002) On the Generalised Affinity Coefficient for Complex Data. *Biocybernetics and Biomedical Engineering,* 22(1), 31-42.

BACELAR-NICOLAU, H., NICOLAU, F.C, SOUSA, Á. & BACELAR-NICOLAU, L. (2009) Measuring Similarity of Complex and Heterogeneous Data in Clustering of Large Data Sets. Biocybernetics and Biomedical Engineering, 29 (2), 9-18.

BACELAR-NICOLAU, H., NICOLAU, F.C., SOUSA, Á., BACELAR-NICOLAU, L (2010) Clustering Complex Heterogeneous Data Using a Probabilistic Approach. *Proceedings of* the *Stochastic Modeling Techniques and Data Analysis International Conference* (*SMTDA2010*), 85-93, (electronic publication).

BOCK, H.-H. & DIDAY, E. (2000) *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data.* Series: Studies in Classification, Data Analysis, and Knowledge Organization, Berlin: Springer-Verlag.

DE CARVALHO, F.A.T., BRITO, P. & BOCK, H.-H. (2006*a*) Dynamic Clustering for Interval Data Based on L_2 Distance. *Computational Statistics,* 21(2), 1-19.

DE CARVALHO, F.A.T., SOUZA, R.M.C.R. de, CHAVENT, M. & LECHEVALLIER, Y. (2006b) Adaptive Hausdorff Distances and Dynamic Clustering of Symbolic Interval Data. *Pattern Recognition Letters,* 27 **(3),** 167-179.

LERMAN, I.C. (1972) *Étude Distributionelle de Statistiques de Proximité entre Structures Algébriques Finies du Même Type: Apllication à la Classification Automatique,* Cahiers du B.U.R.O., 19, Paris.

LERMAN, I.C. (1981) *Classification et Analyse Ordinale des Données,* Paris, Dunod.

NICOLAU, F.C. (1983) Cluster Analysis and Distribution Function. *Methods of Operations Research,* 45, 431-433.

NICOLAU, F.C. & BACELAR-NICOLAU, H. (1998) Some Trends in the Classification of Variables. IN Hayashi, C., Ohsumi, N., Yajima, K., Tanaka, Y., Bock, H.-H., Baba, Y. (Ed.) *Data Science, Classification, and Related Methods.* Springer-Verlag, 89-98.

MATUSITA, K. (1951) On the Theory of Statistical Decision Functions. *Ann. Instit. Stat. Math,* III, 1-30

SOUSA, Á., NICOLAU, F., BACELAR-NICOLAU, H. & SILVA, O. (2013) Clustering of Symbolic Data based on Affinity Coefficient: Application to a Real Data Set. *Biometrical Letters*, 50 (1), 27-38.

SOUZA, R.M.C.R., DE CARVALHO, F.A.T., & PIZZATO, D.F. (2007) A Partitioning Method for Mixed Feature-Type Symbolic Data Using a Squared Euclidean Distance. IN FREKSA, C., KOHLHASE, M., & SCHILL, K. (Ed.) *KI 2006: Advances in Artificial Intelligence.* 29th Annual German Conference on AI, KI 2006, Bremen, Germany, June 14-17, 2006, Proceedings. Series: Lecture Notes in Computer Science, Vol. 4314, Berlin Heidelberg, Springer-Verlag, 260-273.