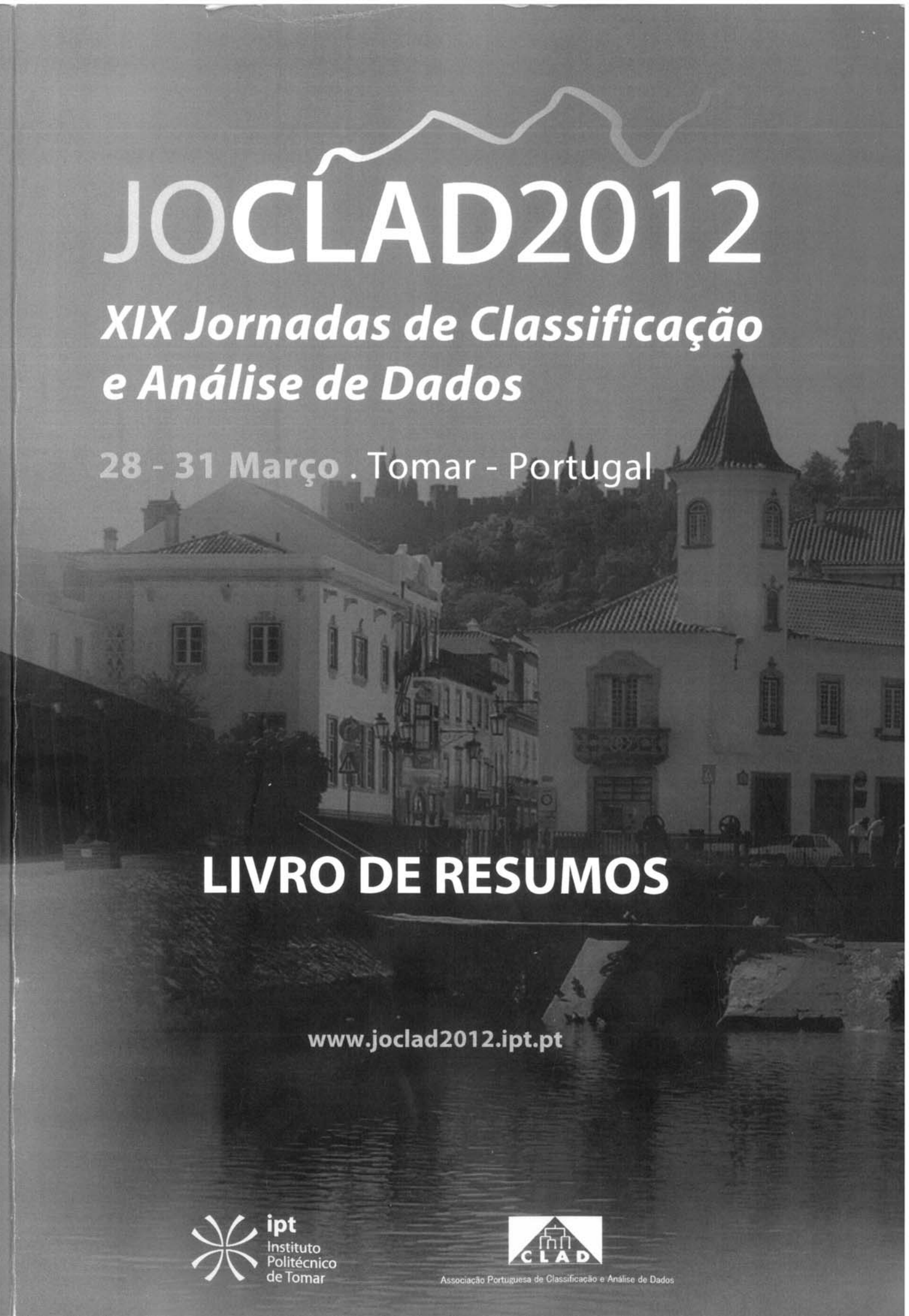

JOCLAD2012

XIX Jornadas de Classificação e Análise de Dados

28 - 31 Março . Tomar - Portugal

LIVRO DE RESUMOS

www.joclad2012.ipt.pt

ipt
Instituto
Politécnico
de Tomar

CLAD
Associação Portuguesa de Classificação e Análise de Dados

S_III.1 Análise Classificatória
6.ª feira, 30 de Março de 2012 (12:30h) – Sala O104

Avaliação e comparação de partições numa perspectiva de qualidade global dos resultados

Osvaldo Silva[1], Helena Bacelar-Nicolau[2], Fernando Nicolau[3]
[1]*Universidade dos Açores, CMATI, osilva@uac.pt;*
[2]*Universidade de Lisboa, Faculdade de Psicologia / Laboratório de Estatística e Análise de Dados, CEAUL, hbacelar@fp.ul.pt;*
[3]*Universidade Nova de Lisboa, FCT, Dep. de Matemática, fnicolau@gmail.com.*

Sumário

Neste trabalho, trata-se de uma metodologia global que tem vindo a ser desenvolvida para avaliar a qualidade dos resultados de uma Análise Classificatória, com base em índices de estabilidade, isolamento e homogeneidade das classes, entre outros. Em complemento, desenvolvemos também um método de visualização, que permite realçar as semelhanças e as diferenças entre as partições e entre os elementos pertencentes a essas partições. A aplicação da metodologia é ilustrada com base num conjunto de dados heterogéneos e de natureza complexa.

Palavras-chave: Análise classificatória hierárquica, coeficiente de afinidade, comparação de partições e metodologia VL.

1 Enquadramento do trabalho

Existem muitos indicadores para a avaliação de classes, partições e classificações hierárquicas, os quais podem diferir substancialmente, quer no tipo de informação utilizada, quer na escala de variação dos seus valores, pelo que se torna desejável sintetizar num único indicador os diferentes índices utilizados, no âmbito de uma metodologia global (Silva et al., 2009, 2010).

No presente trabalho, a avaliação da qualidade global dos resultados de uma Análise Classificatória tem em consideração algumas propriedades importantes, tais como a estabilidade, o isolamento e a homogeneidade das classes. Pretende-se que os índices a utilizar avaliem, de forma global, cada um dos elementos a classificar, cada uma das classes, cada uma das partições e o conjunto de todas as partições em análise.

Seja h um dos m elementos a classificar; c uma das classes de uma partição em k classes; P uma das t partições a avaliar e CP o conjunto de partições. Assumindo que q índices são relevantes para avaliar a qualidade dos resultados de uma Análise Classificatória, pode ser definido um indicador global, variando entre 0 e 1, para avaliar se um elemento/classe/partição/conjunto de partições é melhor, pior ou similar a outro elemento/classe/partição/conjunto de partições. Seja $j=1,...,q$ o índice referente aos índices componentes a utilizar, os quais variam entre 0 e 1, e $U \in \{h, c, P, CP\}$, tendo-se que $U=h$ ou $U=c$

ou *U=P* ou *U=CP*, consoante se esteja a avaliar, respectivamente, um elemento, uma classe, uma partição ou o conjunto das partições. Com base nesta notação, o indicador Global, *Glob_Ind*(*U*), pode ser definido pela seguinte fórmula: $Glob\_Ind(U)=\sum_{j=1}^{q}\alpha_j S_j(U)$, em que os pesos $\alpha_1,\ldots,\alpha_q$ são todos não negativos e a sua soma é igual à unidade, isto é, $\alpha_j \geq 0$, com $j=1,\ldots,q$, e $\sum_{j=1}^{q}\alpha_j = 1$. Assim, o indicador global, *Glob_Ind*(*U*), corresponde a uma combinação linear convexa dos vários índices componentes, $S_j(U)$, com $j=1,\ldots,q$, os quais medem todos uma semelhança/concordância, sendo $S_j(U)$ e q definidos, de forma apropriada, consoante *U=h* ou *U=c* ou *U=P* ou *U=CP*.

Um indicador desta natureza visa comparar as *performances* dos elementos/ classes/partições/conjuntos de partições fornecidas pelos diferentes algoritmos utilizados, usando as informações dos índices considerados mais relevantes, os quais serão aqui designados por índices componentes. Cada um dos índices componentes assume valores que podem, inicialmente, variar em escalas distintas, pelo que devem ser previamente transformados, de modo a que as componentes, $S_j(U)$, com $j=1,\ldots,q$, variem entre 0 e 1, afim de poderem servir de base ao cálculo do indicador global, *Glob_Ind*(*U*).

É, ainda, utilizado um método de visualização, tendo por base somente a informação intrínseca aos dados e aos métodos que vão ser usados aquando da Análise Classificatória, de modo a permitir uma rápida percepção acerca da qualidade dos resultados obtidos ao nível de cada elemento, de cada uma das classes e de cada partição, em cada um dos *t* métodos utilizados (ou *r* reamostragens) e, ainda, do conjunto das partições obtidas (por diferentes métodos ou por reamostragem). Para esse efeito, os valores referentes aos índices *Glob_Ind*(*h*), *Glob_Ind*(*c*), *Glob_Ind*(*P*) e *Glob_Ind*(*CP*) correspondentes, fornecem indicações sobre a qualidade dos resultados obtidos, que facilitam a comparação entre as várias partições obtidas e a escolha, se for caso disso, da partição considerada mais adequada para os dados em análise (Silva et al., 2010).

Foi efectuada a Análise Classificatória Hierárquica de um conjunto de dados heterogéneos e de natureza complexa (Carvalho e Souza, 2009), com base no coeficiente de afinidade generalizado (Nicolau e Bacelar-Nicolau, 1999; Bacelar-Nicolau, 2000, 2002; Bacelar-Nicolau et al. 2009, 2010) e em critérios de agregação clássicos e probabilísticos (Bacelar-Nicolau, 1988; Nicolau, 1983; Nicolau e Bacelar-Nicolau, 1998).

Nesta comunicação, são apresentados os resultados da aplicação da metodologia de avaliação global no que concerne à avaliação das estruturas classificatórias obtidas a partir deste conjunto de dados.

Agradecimentos

Ao Centro de Matemática e Tecnologias de Informação (CMATI) da Universidade dos Açores.

Referências

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*, pp. 181-186. North Holland.

BACELAR-NICOLAU, H. (2000) The Affinity Coefficient. IN BOCK, H.-H. (Ed.) *Analysis of Symbolic Data Exploratory Methods for Extracting Statistical Information from Complex Data*, pp. 160-165. Springer.

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., SOUSA, A. & 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., SOUSA, A. & BACELAR-NICOLAU, L. (2010) Clustering Complex Heterogeneous Data Using a Probabilistic Approach. *Proceedings of Stochastic Modeling Techniques and Data Analysis International Conference (SMTDA2010)*, Chania Crete Greece, 8-11 June 2010 – published on the CD Proceedings of SMTDA2010 (*electronic publication*).

CARVALHO, F. & SOUZA, R. (2009) Unsupervised pattern recognition models for mixed feature-type symbolic data. *Pattern Recognition Letters*, 31 (5), 430-443.

NICOLAU, F. (1983) Cluster Analysis and distribution Function. *Meth. Oper. Res.*, 45, 431-433.

NICOLAU, F. & 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*, pp. 89-98. Springer-Verlag.

NICOLAU, F. & BACELAR-NICOLAU, H. (1999) Clustering Symbolic Objects Associated to Frequency or Probability Laws by the Weighted Affinity Coefficient. IN BACELAR-NICOLAU, H., NICOLAU, F. & JANSSEN, J. (Ed.) *Applied Stochastic Models and Data Analysis. Quantitative Methods in Business and Industry Society*, pp. 155-158. Lisboa, INE.

SILVA, O., BACELAR-NICOLAU, H. & NICOLAU, F. (2009) Como Avaliar a Consistência dos Resultados de uma Análise Classificatória Hierárquica. IN OLIVEIRA, I. et al.

(Eds.) *Actas do XVI Congresso Anual da Sociedade Portuguesa de Estatística*, 2008. Edições S.P.E., 661-672.

SILVA, O., BACELAR-NICOLAU, H. & NICOLAU, F. (2010) Global Approach for Evaluating the Quality of Clustering Results. IN *Programme and Abstracts CFE 10 & ERCIM 10 (4th CSDA International Conference on Computational and Financial Econometrics and 3 rd Conference of the ERCIM Working Group on Computing and Statistics)*, 40.