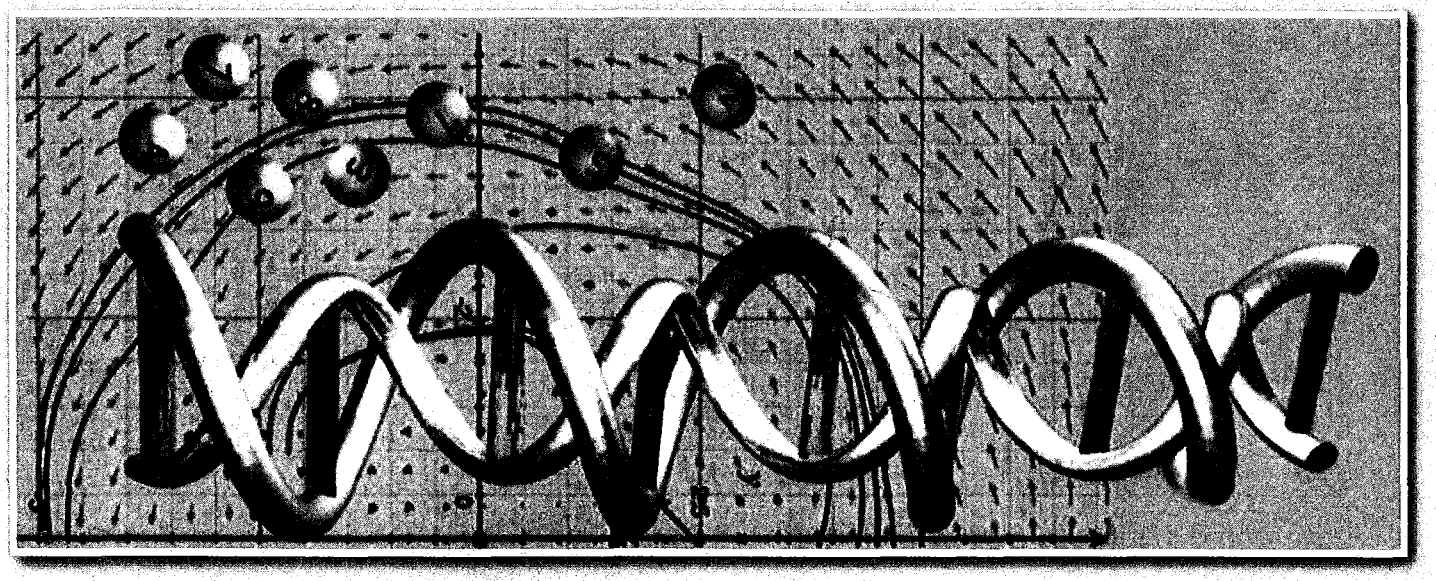
6th WORKSHOP ON STATISTICS, MATHEMATICS AND COMPUTATION

3rd PORTUGUESE - POLISH WORKSHOP ON BIOMETRY



BOOK OF ABSTRACTS

July, 3 - 4 | 2012

Universidade da Beira Interior
COVILHÃ

Chair
Teresa Oliveira (DCeT, Universidade Aberta)
Co-Chair
Stanislaw Mejza (Poznan University of Life Sciences)
Co-Chair
Dário Ferreira (Universidade da Beira Interior)

Organizing Committee
Adérito Marcos (DCeT, Universidade Aberta)
Amílcar Oliveira (DCeT, Universidade Aberta
Célia Nunes (DM, Universidade da Beira Interior)
Dário Ferreira (DM, Universidade da Beira Interior)
Luís Grilo (UDMF, Instituto Politécnico de Tomar)
Rafael Sasportes (DCeT, Universidade Aberta)
Rosário Ramos (DCeT, Universidade Aberta)
Sandra Ferreira (DM, Universidade da Beira Interior)
Teresa Oliveira (DCeT, Universidade Aberta)

Sponsors
Universidade Aberta
Universidade da Beira Interior
Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL)
Centro de Matemática, Universidade da Beira Interior
Poznan University of Life Sciences
Instituto Politécnico de Tomar
Município da Covilhã
PSE - Produtos e Serviços de Estatística
Sumol+Compal
Natura IMB Hotels
ROFF
Museu de Lanifícios, Universidade da Beira Interior
Santander Totta
Turismo Serra da Estrela

Título: Book of Abstracts - 6^{th} Workshop on Statistics, Mathematics and Computation - 3^{rd} Portuguese Polish Workshop on Biometry
Published and printed by: Instituto Politécnico de Tomar
Editors: (Teresa Oliveira, Universidade Aberta), (Stanislaw Mejza, Poznan University of Life Sciences), (Dário Ferreira, Universidade da Beira Interior), (Amílcar Oliveira, Universidade Aberta), (Luís Grilo, Instituto Politécnico de Tomar)
N.º de Depósito Legal: 345933/12
ISBN: 978-972-9473-62-3

WORKSHOP CHAIR

Teresa Oliveira - DCeT, Universidade Aberta

SCIENTIFIC COMMITTEE

Adérito Marcos (DCeT, Universidade Aberta)
Amílcar Oliveira (DCeT, Universidade Aberta)
Ana Maria Portela (Faculdade de Psicologia, Universidade de Lisboa)
Angel A. Juan (Universitat Oberta de Catalunya)
Arminda Manuela Gonçalves (Universidade do Minho)
Áurea Sousa (Universidade dos Açores)
Célia Nunes (Departamento de Matemática, Universidade da Beira Interior)
Dário Ferreira (Departamento de Matemática, Universidade da Beira Interior)
Fátima Brilhante (Universidade dos Açores)
Fernando Branco (Universidade Lusófona)
Fernando Costa (DCeT, Universidade Aberta)
Giovani Silva (Instituto Superior Técnico, Universidade Técnica de Lisboa)
Ivette Gomes (Faculdade de Ciências da Universidade de Lisboa)
Iwona Mejza (Faculty Agronomy Bioeng., Poznan University of Life Sciences, Poland)
Jaime Remédios (DCeT, Universidade Aberta)
Javier Martín Vallejo (Facultad de Medicina, Universidad de Salamanca)
João Araújo (DCeT, Universidade Aberta)
Luís Grilo (Instituto Politécnico de Tomar)
Manuela Neves (ISA, Universidade Técnica de Lisboa)
Manuela Oliveira (Departamento de Matemática, Universidade de Évora)
Maria Antónia Huertas (Universitat Oberta de Catalunya, Barcelona, Spain)
Maria Antónia Turkman (DEIO, Faculdade de Ciências, Universidade de Lisboa)
Maria do Rosário Almeida (DCSH, Universidade Aberta)
Maria José Fernández Gómez (Facultad de Medicina, Universidad de Salamanca)
Marina Andrade (ISCTE, Instituto Universitário de Lisboa)
Miguel Pardal (Department of Life Sciences, University of Coimbra)
Purificación Galindo (Facultad de Medicina, Universidad de Salamanca)
Rafael Sasportes (DCeT, Universidade Aberta)
Rosário Ramos (DCeT, Universidade Aberta)
Sandra Ferreira (Departamento de Matemática da Universidade da Beira Interior)
Sónia Seixas (DCeT, Universidade Aberta)
Stergios Tzortzios (Faculty of Agricultural Sciences, University of Thessaly, Greece)
Teresa Oliveira (DCeT, Universidade Aberta)
Ulisses Azeiteiro (DCeT, Universidade Aberta)
Valter Vairinhos (ISLA Santarém)
Vítor Rocio (DCeT, Universidade Aberta)

Clustering of Symbolic Data based on Affinity Coefficient: Application to real data sets

Áurea Sousa
University of Azores, Department of Mathematics, CEEAplA, CMATI, Portugal, aurea@uac.pt

Helena Bacelar-Nicolau
University of Lisbon, Faculty of Psychology, Laboratory of Statistics and Data Analysis, Portugal, hbacelar@fp.ul.pt

Fernando C. Nicolau
New University of Lisbon, FCT, Department of Mathematics, Portugal, geral@datascience.org

Osvaldo Silva
University of Azores, Department of Mathematics, CMATI, Portugal, osilva@uac.pt

Abstract: The increasing use of databases, often large ones, in diverse areas of study makes it pertinent to summarise data in terms of their most relevant concepts. These concepts may be described by types of complex data, also known as symbolic data. In a symbolic data table, lines correspond to symbolic objects (SO's) and columns to symbolic variables, which may contain not just one value, as usual, but values, such as subsets of categories, intervals in real axes, or frequency distributions. Symbolic data arise in a number of different ways (for example, as the result of aggregation of large data sets to obtain a data set of manageable size, or as a result of some scientific question(s) of interest). The aim of Symbolic Data Analysis (SDA) is to extend classical data analysis techniques to these kinds of data.
Based on the affinity coefficient between two discrete probability distributions as defined by the pioneer work of Matusita, started in 1951, Bacelar-Nicolau suggested the use of the affinity coefficient as a basic similarity coefficient between the columns or the lines of a data matrix. Later on she extended that coefficient to different types of data, including complex data (symbolic data) and variables of mixed types (heterogeneous data), possibly with different weights.
We present some results from the Ascendant Hierarchical Cluster Analysis (AHCA) of symbolic objects described by interval data, in order to illustrate the effectiveness of the Ascendent Hierarchical Cluster Analysis based on the weighted generalized affinity coefficient, for symbolic data. The measure of comparison between the elements was combined with classical aggregation criteria and probabilistic ones. The probabilistic aggregation criteria used in this study belong to a parametric family of methods in the scope of the probabilistic approach of AHCA, named VL methodology and the validation of the clustering results is based on some validation measures. Finally, we compare the results achieved by our approach with the ones obtained by other authors.

Keywords: Ascendant Hierarchical Cluster Analysis, Symbolic Data, Interval Data, Affinity Coefficient, VL Methodology.

References

Bacelar-Nicolau, H. (1988). Two Probabilistic Models for Classification of Variables in Frequency Tables. In: Classification and Related Methods of Data Analysis, H. -H. Bock (ed.), North Holland: Elsevier Publishers B.V, 181-186.

Bacelar-Nicolau, H. (2000). The Affinity Coefficient for Complex Data. In: Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data, H. -H. Bock, E. Diday (Eds.), 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, v. 22, n.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, v. 29, n. 2, 9-18, 2009.

Bock, H. -H., Diday, E. (Eds.) (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.

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

Nicolau, F. (1980). Critérios de Análise Classificatória Hierárquica baseados na Função de Distribuição. Tese de Doutoramento, FCL, Universidade de Lisboa.

Nicolau, F., Bacelar-Nicolau, H. (1998). Some Trends in the Classification of Variables. In: Data Science, Classification, and Related Methods, C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H. -H. Bock, Y. Baba (Eds.), Springer-Verlag, 89-98.

Nicolau, F., Bacelar-Nicolau, H. (1999). Clustering Symbolic Objects Associated to Frequency or Probability Laws by the Weighted Affinity Coefficient. In: Applied Stochastic Models and Data Analysis. Quantitative Methods in Business and Industry Society, H. Bacelar-Nicolau, F. C. Nicolau and Jacques Janssen (Eds.), INE, Lisbon, Portugal, 155-158.

Sousa, A. (2005). Contribuições à Metodologia VL e Índices de Validação para Dados de Natureza Complexa. Tese de Doutoramento, Universidade dos Açores, Ponta Delgada.

Sousa, A., Nicolau, F., Bacelar-Nicolau, H., Silva, O. (2010). Weighted Generalised Affinity Coefficient in Cluster Analysis of Complex Data of the Interval Type. Biometrical Letters, v. 47, n. 1, 45-56.