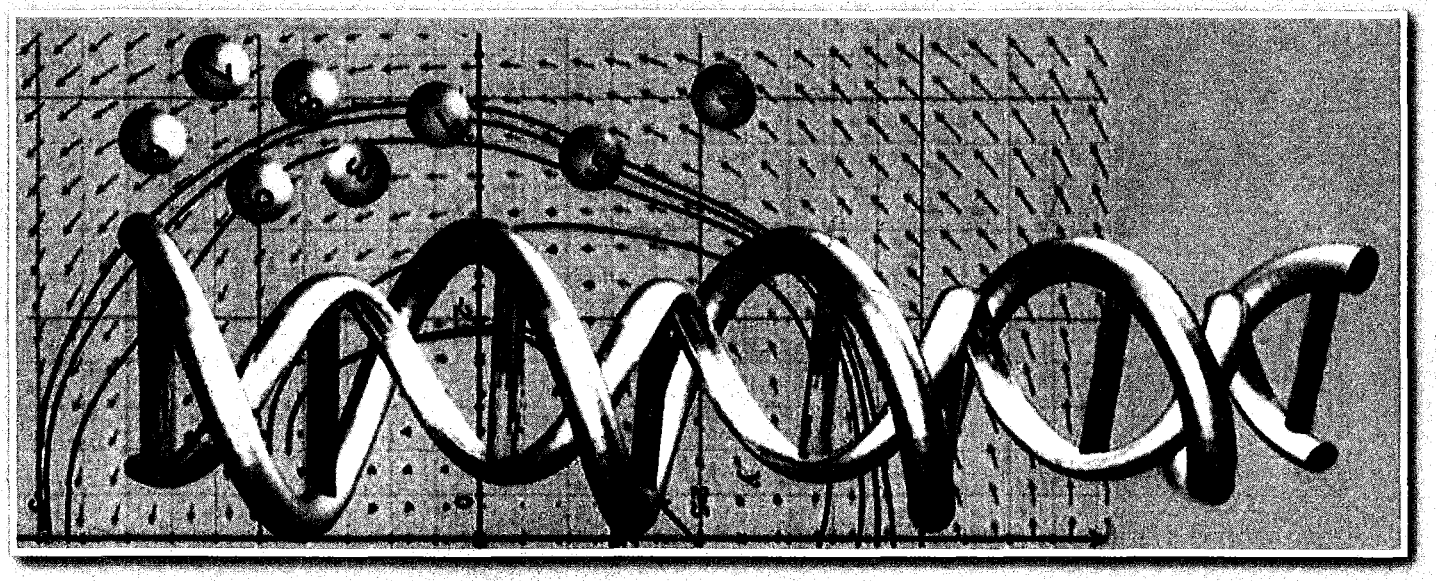
6th WORKSHOP ON STATISTICS, MATHEMATICS AND COMPUTATION

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

Global Approach for the comparison of Clustering Results

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

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

Fernando Nicolau
New University of Lisbon, Faculty of Science and Technology, Portugal, geral@datascience.org

Abstract: The extraction of useful knowledge from a Hierarchical Cluster Analysis (HCA) is a complex process which depends on many factors, such as the applied clustering algorithms and the strategies developed in the initial stage of the HCA. We present a global approach for evaluating the quality of clustering results based on the comparison of partitions from the different clustering algorithms using the most relevant information available (e.g. stability, isolation and homogeneity of the clusters). In addition, we suggest a visual method to facilitate the evaluation of the quality of the partitions that allows us a quick perception of the similarities and the differences between the partitions, including the behaviour of the elements in the partitions. We illustrate our approach using a real data set (horse data). We applied HCA based on the weighted generalized affinity coefficient (similarity coefficient) to the case of complex data (symbolic data), combined with 26 clustering (classic and probabilistic) algorithms. Finally, we discuss the obtained results and the contribution of this approach to a better knowledge on the cluster structure of a data set.
Keywords: Cluster Analysis, VL Methodology, Affinity Coefficient, Comparing Partitions, Cluster Stability and Cluster Validation.

References

Bacelar-Nicolau, H. (2000). The Affinity Coefficient. In: Analysis of Symbolic Data Exploratory Methods for Extracting Statistical Information from Complex Data, H.H. Bock, E. Diday (Eds.), Springer, 160-165.
Bacelar-Nicolau, H. (2002). On the Generalised Affinity Coefficient for Complex Data. Biocybernetics and Biomedical Engineering, Vol. 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, Vol. 29, n° 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, Volume 31, n°5, 430-443.
Gordon, A D. (1999). Classification, 2nd.Chapman & Hall, London.
Lerman, I. C. (1981). Classification et Analyse Ordinale des Données. Paris, Dunod.
Nicolau, F. (1983). Cluster Analysis and distribution Function in Meth. Oper. Res., 45, 431-433.
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. Nicolau and Jacques Janssen (Eds.), INE, Lisboa, Portugal, 155-158.
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 3rd Conference of the ERCIM Working Group on Computing and Statistics), 40.
Silva, O. (2012). Contributos para a Avaliação e Comparação de Partições em Análise Classificatória. Tese de Doutoramento, Universidade dos Açores, Ponta Delgada.