MEASURING SIMILARITY OF COMPLEX AND HETEROGENEOUS DATA IN CLUSTERING OF LARGE DATA SETS

Helena Bacelar-Nicolau[1], Fernando Nicolau[2], A.Urea Sousa[3], Leonor Bacelar-Nicolau[4]

[1]*Universidade de Lisboa. FPCE, Laboratório de Estatistica e Analise de Dados, Lisboa, Portugal*

[2] *Universidade Nova de Lisboa, FCT, Departamento de Matematica, Caparica, Portugal*

[3] *Universidade dos Açores, Departamento de Matematica, Ponta Delgada, Açores, Portugal*

[4] *Universidade de Lisboa. FML, Laboratório de Biomatematica & Instituto de Medicina Preventiva. Lisboa, Portugal*

Abstract:

Cluster analysis or classification usually concerns a set of exploratory multivariate data analysis methods and techniques for finding a clustering structure on a dataset. That may refer either to groups of statistical data units or to groups of variables. In this work we deal with a generalization of this paradigm concerning clustering of complex data described by three different types of variables, frequent1y present in a three-way context. We obtain compatible versions of the same affinity coefficient for measuring similarity between statistical data units described by those three types of variables. A global generalized similarity coefficient is analyzed for such kind of mixed data, often arising in data mining or knowledge mining.

Keywords: cluster analysis, different type variables, similarity coefficient, three-way data