Normalizing Constrained Symbolic Data for Clustering

Marc Csernel*, F.A.T. de Carvalho **

*INRIA - Rocquencourt, Domaine de Voluceau
Rocquencourt - B. P. 105, 78153 le Chesnay Cedex - France
Marc.Csernel@inria.fr

**Centro de Informatica - CIn/UFPE, Av. Prof Luiz Freire,
s/n,Cidade Universitaria, CEP 50.740-540, Recife - PE BRAZIL
fatc@cin.ufpe.br

Abstract. Clustering is one of the most common operation in data analysis while constrained is not so common. We present here a clustering method in the framework of Symbolic Data Analysis (S.D.A) which allows to cluster Symbolic Data. Such data can be constrained relations between the variables, expressed by rules which express the domain knowledge. But such rules can induce a combinatorial increase of the computation time according to the number of rules. We present in this paper a way to cluster such data in a quadratic time. This method is based first on the decomposition of the data according to the rules, then we can apply to the data a clustering algorithm based on dissimilarities.

1 Introduction.

The aim of cluster analysis is to organize a set of items into clusters such that items within a given cluster have a high degree of similarity, whereas items belonging to different clusters have a high degree of dissimilarity. Cluster analysis can be divided into hierarchical and partitioning methods (Gordon (1999), Everitt (2001)). While hierarchical methods build hierarchies, i.e., a nested sequence of partitions of the input data, partitioning methods try to obtain a partition of the input data into a fixed number of clusters, usually by optimizing a function.

This paper addresses the partitioning of constrained symbolic data into a predefined number of clusters. Symbolic data allows to manage some domain knowledge, provided by relations between the variables. These relations are expressed by rules expressing knowledge among the data. A good description of symbolic data can be found in Bock and Diday (2000).

Symbolic data are expressed by symbolic variables which are defined according to the type of their domain. In this paper we will focus on set-valued variables which take their values in a set of nominal categories and a list-valued variable which take as values list of ordered categories.

Table 1 displays an example of two symbolic descriptions called d_1 and d_2, which are described by three set-valued variables and one list-valued variable (Thorax size). These data can be constrained by the dependencies rules r_1, r_2.