From complex unstructured heterogeneous massive data to Symbolic Data Analysis

Edwin Diday
CEREMADE, University Paris-Dauphine, Paris, France diday8@ceremade.dauphine.fr

Many application fields generate complex, unstructured, heterogeneous and massive amounts of data that are difficult to analyze with traditional techniques. In recent years the Symbolic Data Analysis (SDA) framework has developed a theory and practice addressed to such kind of data. The fusion of such complex data leads to "symbolic data tables", where the units are categories or classes of standard samples, described by intervals, distributions, sets of categories, and the like. SDA becomes a challenging task in extending standard Data Analysis and Data Mining to such data. In this talk we propose a strategy for extending standard principal component analysis (PCA) to such data in the case where the variables values are "bar charts". The results are illustrated by two applications. The first is industrial and concerns Power Plant insecurity, the second is in epidemiology and concerns trachoma study.

Keywords: principal component analysis; symbolic data analysis; copulas; compositional data; metabins.