The symbolic data analysis paradigm, discriminant discretization and financial application

Edwin Diday**, Filipe Afonso*, Raja Haddad*,***

*Syrokko, Aéropôle Roissy CDG, 95731 Roissy
(haddad, afonso)@syrokko.com
(**CEREMADE, ***LAMSADE), Université de Paris 9 Dauphine,75775 Paris
diday@ceremade.dauphine.fr

Abstract. The variability inside classes of individuals, categories (defined by a categorical variable) or concepts (defined by an intent and an extent, like species for example), is expressed by the use of intervals, histograms, distributions, sequences of weighted values and the like. In this way we obtain new kinds of data called "symbolic". The aim of "Symbolic Data Analysis" (SDA) is to study and extract new knowledge from these new kinds of data by an extension of Statistics and Data Mining to symbolic data. We show that SDA is a new paradigm opened to a vast field of research and applications. Then, we give a way for obtaining discriminate symbolic descriptions by an original discretisation method, which is illustrated by a financial application.

1 Introduction

The usual data mining model is based on two parts: the first concerns the observations (i.e., observed entities), the second contains their description by several standard numerical or categorical variables. The Symbolic Data Analysis (SDA) model (see (Diday, 1987), (Billard and Diday, 2006), (Diday and Noirhomme, 2008)) needs two more parts: the first concerns higher level units defined by classes of observations, categories (i.e. a name given to a class) or concepts (defined by an intent and an extent) and the second concerns their description by "symbolic data" which may be standard categorical or numerical data but moreover intervals, histograms, sequences of weighted values and the like, in order to take care of the variability of the observations inside each class. These new kinds of data are called "symbolic" as they cannot all be manipulated like numbers.

Based on this model, new knowledge can be extracted by new tools of data mining extended to this higher level units considered as new kinds of observations. Inspired by "The Structure of Scientific Revolutions" (Kuhn, 1962), the second section of this paper tries to show that the "Symbolic Data Analysis" framework is a new scientific paradigm by answering the following questions: what is the failure in the actual practice? What is the paradigm shift? What is to be observed and scrutinized? What kind of questions are there, and how are they structured? What are the principles and the theoretical development? What is the applicability domain?

In order to build symbolic data from an initial standard data table, a discretization process is needed and constitutes a fundamental part of the aggregation process which leads to histogram