COMMENT EXTRAIRE DES CONNAISSANCES A PARTIR DES CONCEPTS DE VOS BASES DE DONNEES ?

LES DEUX ETAPES DE L'ANALYSE DES DONNEES SYMBOLIQUES.

E. Diday

Université Paris-Dauphine

diday@ceremade.dauphine.fr

Résumé

Vos bases de données contiennent des concepts sous-jacents. Ils sont associés aux catégories issues de produits cartésiens de variables qualitatives ou de classifications automatiques. Ces concepts constituent alors des unités d'étude d'un niveau de généralité supérieur aux données initiales. Ce niveau est souvent désiré par les utilisateurs mais freinés par le carcan des données classiques qui ne tiennent pas compte de la variation des instances de ces concepts. L'analyse des données symboliques (ADS) a pour objectif dans une première étape de constituer ces concepts et de les décrire en prenant en compte leur variation interne par des variables dites « symboliques » (à valeur intervalle, histogramme, lois etc.) car non manipulables comme des nombres. La seconde étape d'une ADS consiste à les analyser. Pour cela on est amené à étendre les méthodes de la statistique exploratoire et de la fouille de données aux données symboliques (ces méthodes deviennent alors des cas particuliers d'ADS) et de développer des outils nouveaux spécifiques. On montre que ces données ne peuvent pas être réduites à des données classiques. On décrit les quatre espaces de la modélisation sous-jacente où les concepts sont modélisés par des objets symboliques, puis la modélisation mathématique des données (sous forme de variables à valeur variable aléatoire) et des classes ainsi que de leur structure en généralisant les treillis de Galois, hiérarchies, pyramides classiques aux données symboliques. On introduit leur classification spatiale étendant les cartes de Kohonen à des données et des structures pyramidales plus riches. On termine enfin par une application industrielle et la présentation du logiciel SODAS issu de deux projets européens d'EUROSTAT.

Mots clés : Data Mining, fouille de données, analyse des données, statistique descriptive, analyse des données exploratoire, données symboliques, classification automatique, analyse factorielle, treillis de Galois stochastiques, pyramides, analyse de concepts, classification spatiale.