Principes d'Analyse des données symboliques et application à la détection d'anomalies sur des ouvrages publics

E. Diday *, C. Crémona**, F. Goupil*, F. Afonso***, M. Rahal*

*CEREMADE – Paris Dauphine, Place du Mal de Lattre de Tassigny 75775 Paris Cedex 16
(goupil, diday, rahal) @ceremade.dauphine.fr
** Laboratoire Central des Ponts et Chaussées. 75006 Paris.
***SYROKKO - 6, rue Ambroise Jacquin - 95190 Fontenay-en-Parisis
afonso@syrokko.com

Résumé. L'analyse des données Symboliques a pour objectif de fournir des résultats complémentaires à ceux fournis par la fouille de données classique en créant des concepts issus de données simples ou complexes puis en analysant ces concepts par des descriptions symboliques où les variables expriment la variation des instances de ces concepts en prenant des valeurs intervalle, histogramme, suites, munies de règles et de taxonomies, etc.

1 Introduction

On appelle « concept », une entité qui se définit par un croisement de catégories. L'objet de l'ADS est d'analyser des ensembles de concepts décrits par des variables symboliques. Ces variables sont non seulement à valeur numérique ou qualitative mais aussi à valeur intervalle, histogramme, loi de probabilité, fonction, ensemble de valeurs etc., afin de tenir compte de la variation des valeurs prises par les individus de l'extension de chaque concept. L'ADS et son logiciel SODAS comportent deux étapes : la première consiste à construire la description des concepts à partir de celle des individus, la seconde consiste à analyser le tableau de données symboliques ainsi créé en étendant les méthodes de la Statistique ou du Data Mining aux concepts considérés comme unités statistiques de plus haut niveau. Nous illustrons ces deux étapes en montrant trois avantages de l'ADS : i) on peut étudier les bonnes unités statistiques à un niveau de généralisation voulu par l'utilisateur ; ii) on réduit la taille des données en considérant comme unités d'étude, des classes plutôt que les individus ; iii) on réduit le nombre de variables du fait qu'elles sont à valeur symbolique (par exemple, à valeur « histogramme » plutôt qu'à valeur «fréquence d'une catégorie» ou à valeur intervalle plutôt qu'à valeur « borne d'intervalle »). On utilise pour cela le logiciel SODAS (voir l'ouvrage collectif issu du projet européen ASSO d' EUROSTAT : Diday, Noirhomme (2007)).

2 Description

Les données fournies par le LCPC (Laboratoire Central des Ponts et Chaussées) sont constituées d'un ensemble de 14 TGV qui en passant à une température donnée sur un pont déclenchent des signaux de 9 capteurs répartis à différents endroits du pont (voir la figure 1). En entrée, on dispose d'un tableau de données symboliques qui contient dans la case (i, j) le