Un aperçu de la fouille visuelle de données

Hanene Azzag*, David Da Costa**
Christiane Guinot***, **, Gilles Venturini**

*Laboratoire d'Informatique de Paris-Nord
99 Avenue J-B. Clément, 93430 Villetaneuse.
hanene.azzag@lipn.univ-paris13.fr
**Laboratoire d'Informatique, Université François-Rabelais de Tours,
64, Avenue Jean Portalis, 37200 Tours.
david.dacosta@univ-tours.fr, venturini@univ-tours.fr
***CE.R.I.E.S., 20 rue Victor Noir, 92521 Neuilly-sur-Seine Cedex
christiane.guinot@ceries-lab.com

Résumé. Nous présentons dans cet article un aperçu de la fouille visuelle de données. Pour commencer, nous situons ce domaine par rapport à d'autres approches et nous en rappelons les principes fondateurs. Ensuite, nous montrons qu'il existe de nombreux points de vue pour aborder les travaux en fouille visuelle de données : les données ou connaissances à visualiser, la tâche à accomplir, la représentation visuelle choisie, la méthode de calcul de cette représentation ou encore le domaine d'application traité. Nous choisissons tout d'abord le point de vue des données à visualiser en détaillant des approches représentatives pour la visualisation de données numériques, de données hiérarchiques et de documents. Ensuite, notre prenons le point de vue de la représentation visuelle choisie en présentant le domaine des métaphores visuelles utilisées pour la fouille de données. Nous finissons en traitant d'un domaine thématique particulier, l'analyse d'audience d'un site Web, et en concluant sur les perspectives en fouille visuelle de données.

1. Introduction

La base des Iris de Fisher (Fisher 1936) est un ensemble de données bien connu maintenant (Blake et Merz 1998) composé de 150 fleurs décrites par 4 caractéristiques numériques et un attribut de classes (3 classes possible). Imaginons que l'expert du domaine souhaite comprendre comment cette base est structurée et en particulier s'il existe du bruit, des groupes de points distincts ayant certaines caractéristiques, etc. Une première approche possible pour résoudre ce type de problème consiste à utiliser des outils d'apprentissage artificiel qui vont extraire par exemple pour chaque classe, des domaines de valeurs qui font que chaque groupe appartient à telle ou telle classe. Un avantage incontestable de ce type de méthode, si l'on omet les étapes souvent cruciales de réglages des paramètres, est d'être entièrement automatique et de ne pas faire appel à l'expert du domaine dans le processus de découverte des connaissances. Cependant, on peut aussi donner des inconvénients de ce type de méthodes : d'une part la compréhensibilité des résultats n'est pas toujours évidente pour l'expert, que cela soit en terme de représentation des hypothèses apprises (langage de représentation nécessitant quel investissement de la part de l'expert) ou d'explications sur la