Prétraitement de grands ensembles de données pour la fouille visuelle

Edwige Fangseu Badjio, François Poulet

ESIEA Pôle ECD,
Parc Universitaire de Laval-Changé,
38, Rue des Docteurs Calmette et Guérin,
53000 Laval France
fangseubadjio@esiea-ouest.fr
poulet@esiea-ouest.fr

Résumé. Nous présentons une nouvelle approche pour le traitement des ensembles de données de très grande taille en fouille visuelle de données. Les limites de l'approche visuelle concernant le nombre d'individus et le nombre de dimensions sont connues de tous. Pour pouvoir traiter des ensembles de données de grande taille, une solution possible est d'effectuer un prétraitement de l'ensemble de données avant d'appliquer l'algorithme interactif de fouille visuelle. Pour ce faire, nous utilisons la théorie du consensus (avec une affectation visuelle des poids). Nous évaluons les performances de notre nouvelle approche sur des ensembles de données de l'UCI et du Kent Ridge Bio Medical Dataset Repository.

1 Introduction

Nous nous intéressons au problème de prétraitement de grands ensembles de données. Notre but est de réduire les informations contenues dans les ensembles de données volumineux aux informations les plus significatives. Il existe des techniques expérimentalement validées pour ce faire. D'un point de vue applicatif, un problème majeur se pose quant au choix d'une de ses méthodes. Une solution qui constitue notre contribution dans ce travail serait d'utiliser une combinaison de techniques ou de stratégies. A cet effet, nous nous appuyons sur la théorie du consensus. L'utilisation de cette combinaison de stratégies ou d'expertises peut être justifiée par l'un des faits suivants :

- il n'est pas possible de déterminer a priori quelle méthode de sélection de sous-ensemble d'attributs est meilleure que toutes les autres (en tenant compte des différences entre le temps d'exécution et la complexité),
- un sous-ensemble optimal d'attributs n'est pas nécessairement unique,
- la décision d'un comité d'experts est généralement meilleure que la décision d'un seul expert.

Les résultats obtenus après des expérimentations permettent de conclure que l'approche proposée réduit de façon significative l'ensemble de données à traiter et permet de les traiter interactivement. Cette contribution commence par un état de l'art et la problématique du