Algorithme semi-interactif pour la sélection de dimensions

Lydia Boudjeloud, François Poulet

ESIEA Pôle ECD
38, rue des docteurs Calmette et Guérin
Parc Universitaire de Laval-Changé
53000 Laval
boudjeloud|poulet@esiea-ouest.fr

Résumé. Nous présentons un algorithme génétique semi-interactif de sélection de dimensions dans les grands ensembles de données pour la détection d'individus atypiques (outliers). Les ensembles de données possédant un nombre élevé de dimensions posent de nombreux problèmes aux algorithmes de fouille de données, une solution est d'effectuer un pré-traitement afin de ne retenir que les dimensions "intéressantes". Nous utilisons un algorithme génétique pour le choix du sous-ensemble de dimensions à retenir. Par ailleurs nous souhaitons donner un rôle plus important à l'utilisateur dans le processus de fouille, nous avons donc développé un algorithme génétique semi-interactif où l'évaluation des solutions n'élimine pas complètement la fonction d'évaluation mais la couple avec une évaluation de l'utilisateur. Enfin, l'importante réduction du nombre de dimensions nous permet de visualiser les résultats de l'algorithme de détection d'outlier. Cette visualisation permet à l'expert des données d'étiqueter les éléments atypiques (erreurs ou simplement des individus différents de la masse).

1 Introduction

Nous nous intéressons à la recherche d'outliers (individus atypiques) dans les ensembles de données ayant un grand nombre de dimensions. Pour pouvoir traiter de tels ensembles de données (par exemple les ensembles de données de fouille de texte ou de bio-informatique), la plupart des algorithmes de fouille de données actuels nécessitent un prétraitement permettant de réduire le nombre de dimensions (avec plus ou moins de perte d'information). L'approche la plus intuitive pour appréhender le problème des grandes dimensions est d'énumérer tous les sous-ensembles de dimensions possibles et de rechercher le sous-ensemble qui satisfait la problématique traitée. Cependant, le fait d'énumérer (rechercher) toutes les combinaisons possibles est un problème NP-difficile (Narenda et Fukunaga, 1977). Parmi les solutions proposées pour ce problème, on retrouve la réduction de dimensions (combinaison de dimensions, généralement linéaire) et la sélection de dimensions (on n'utilise qu'un sous-ensemble des dimensions originales). L'avantage de cette dernière solution est que nous ne perdons pas l'information que pourrait apporter la dimension, car elle est considérée individuellement non en combinaison (linéaire) avec d'autres dimensions. Les techniques de sélection de dimensions consistent donc à réduire l'ensemble des