Un survol des algorithmes évolutionnaires dans la fouille de données

Fatima-Zohra Kettaf*
Jean-Pierre Asselin de Beauville**

*IRIT-UPS, UMR 5505, 118 route de Narbonne 31062 toulouse cedex 04 France
kettaf@irit.fr
**Laboratoire d'Informatique, Ecole Polytechnique, Université de Tours, 64 avenue Jean Portalis 37200 Tours, Détaché Temporaire à l'Agence Universitaire de la Francophonie au Canada (Montréal)
jean-pierre.asselin@auf.org

Résumé. Cet article a pour objectif la présentation des travaux récents dans le domaine de la fouille de données, basés sur l'évolution génétique.

1 Introduction

La fouille de données peut se définir comme un ensemble de méthodes permettant d'analyser des données déjà collectées dans de très grandes bases de données (transactions bancaires, séquences biologiques, ...) afin d'extraire des relations ou des structures ayant une sémantique utile pour les utilisateurs.
Les algorithmes évolutionnaires (AE), par leur capacité d'exploration de grands espaces de solutions, se sont révélés être des outils utiles et efficaces dans le processus de fouille de données. On aborde, dans cet article, leur contribution aux phases : de pré-traitement, de découverte de règles et de post-traitement.

2 La fouille de données et le processus de découverte de connaissances

La fouille de données [?] peut se diviser en trois étapes : le pré-traitement, l'extraction de concepts, et le post-traitement. Les résultats sont généralement des règles permettant d'expliquer les données et/ou de faire des prédictions. Les entrées du processus sont souvent des données brutes qui vont subir des traitements de nature différente allant des procédures de bas niveau telles que la discrétisation et le filtrage, aux procédures de haut niveau, telles que l'extraction de concepts et de règles. Pour une application donnée, il existe plusieurs profils d'utilisateur en fonction des objectifs fixés. Ces objectifs sont généralement regroupés en cinq catégories : exploratoires, descriptifs, portés sur l'analyse de dépendances, prédictifs ou encore tournés vers la recherche par contenu. Pour l'exploration, les techniques associées font appel à l'interaction et à la visualisation. On utilise souvent l'analyse en composantes principales (ACP) pour réduire la dimension de l'espace de représentation et faciliter la visualisation. Pour ce qui concerne la description, il est souvent utile de supposer des modèles probabilistes,