Détection de séquences atypiques basée sur un modèle de Markov d'ordre variable

Cécile Low-Kam*, Anne Laurent**
Maguelonne Teisseire**

*I3M, Univ. Montpellier 2 - CNRS, Pl. Eugène Bataillon, Montpellier, France
cecile.lowkam@math.univ-montp2.fr
**LIRMM, Univ. Montpellier 2 - CNRS, 161, rue Ada, Montpellier, France
{laurent,teisseire}@lirmm.fr

Résumé. Récemment, le nombre et le volume des bases de données séquentielles biologiques ont augmenté de manière considérable. Dans ce contexte, l'identification des anomalies est essentielle. La plupart des approches pour les extraire se fondent sur une base d'apprentissage ne contenant pas d'outlier. Or, dans de très nombreuses applications, les experts ne disposent pas d'une telle base. De plus, les méthodes existantes demeurent exigeantes en mémoire, ce qui les rend souvent impossibles à utiliser. Nous présentons dans cet article une nouvelle approche, basée sur un modèle de Markov d'ordre variable et sur une mesure de similarité entre objets séquentiels. Nous ajoutons aux méthodes existantes un critère d'élagage pour contrôler la taille de l'espace de recherche et sa qualité, ainsi qu'une inégalité de concentration précise pour la mesure de similarité, conduisant à une meilleure détection des outliers. Nous démontrons expérimentalement la validité de notre approche.

1 Introduction

Un outlier est défini dans (Hawkins (1980)) comme "*une observation qui s'écarte tellement des autres qu'elle est susceptible d'avoir été générée par un mécanisme différent*". Ces dernières années, la détection d'outliers a été étudiée pour des types de données très divers.

En effet, les applications associées à la découverte d'anomalies sont très nombreuses, dans des domaines aussi variés que la détection de fraudes ou l'analyse de séquences biologiques. Parmi elles, les bases d'ADN et de protéines ont fait l'objet de nombreuses études pour une meilleure compréhension des phénomènes biologiques, par exemple par l'extraction de motifs (Ferreira et Azevedo (2007)). La perspective d'identifier des anomalies peut alors compléter les propositions actuelles.

Mais effectuer cette recherche demeure problématique, puisque les outliers sont rares par définition. De plus, ils ne doivent pas être confondus avec le bruit inhérent à tout jeu de données. Néanmoins, certaines propositions existent et nous pouvons citer celles fondées par exemple sur des tests de discordance, sous l'hypothèse d'une distribution de probabilités des observations donnée, dans le cadre univarié ou multivarié (Barnett et Lewis (1994)). D'autres