Affectation pondérée par le critère de Kolmogorov-Smirnov sur des données de type intervalle et diagramme

Chérif Mballo
Laboratoire de bioinformatique, Département d'informatique
Université du Québec à Montréal, Case Postale 8888
Succursale Centre Ville, Montréal (QC) H3C 3P8 Canada
Courriel : mballo.cherif@courrier.uqam.ca

Résumé. Le critère de découpage binaire de Kolmogorov-Smirnov a été introduit par (Friedman, 1977) pour une partition binaire à expliquer sur des variables continues. Nos travaux antérieurs nous ont permis de l'étendre dans le cas où les objets destinés à être classés par un arbre de décision sont décrits par des variables de type intervalle et diagramme ((Mballo et Diday, 2004), (Mballo et al., 2004)) en adoptant une affectation pure. Dans cet article, nous proposons une méthode permettant d'affecter une donnée à la fois aux deux nœuds fils générés par le partitionnement d'un nœud non terminal. Cette approche d'affectation est basée sur des poids et tient compte de la position de la donnée à classer par rapport à celle seuil de coupure.

1 Introduction

Avec l'avènement de l'analyse des données symboliques (Bock et Diday, 2000), on assiste à la mise au point de méthodes de construction d'arbres de décision sur des données de type intervalle et diagramme ((Périnel, 1996), (Aboa, 2002), (Vrac, 2002), (Limam, 2005)). Pour construire l'arbre de décision, ces auteurs utilisent l'entropie, le critère de Gini, le gain ratio et le likelihood comme critère d'évaluation de la qualité d'une coupure.

Dans cet article, nous nous intéressons au critère de découpage binaire de Kolmogorov-Smirnov, noté KS dans la suite. Ce critère a été introduit par (Friedman, 1977) pour une partition binaire à expliquer sur des variables continues. Il a été également exploré quelques années plutard par (Utgoff et Clouse, 1996) sur ce même type de données. (Asseraf, 1998) s'est intéressé à son extension aux données qualitatives. Il présente un bon pouvoir discriminant sur des données classiques. Dans ((Mballo et Diday, 2004), (Mballo et al., 2004)), nous l'avons étendu aux données de type intervalle et diagramme mais dans cette approche, une donnée est entièrement affectée à un nœud (affectation pure). Comme ce critère nécessite un ordre des données, nous présentons tout d'abord quelques méthodes pour ordonner des intervalles (Diday et al., 2003) et des diagramme. La possibilité d'estimer la fonction de répartition théorique par celle empirique nous permet d'adapter ce critère aux données de type intervalle et diagramme. Nous présentons à la section 4 une méthode permettant d'affecter une donnée à la fois aux deux nœuds fils générés par le partitionnement d'un nœud non terminal. La motivation de cette approche d'affectation est de prendre en compte le positionnement de la donnée à classer par rapport à la donnée seuil de coupure en définissant des poids. Des exemples illustrant cette approche d'affectation sont également présentés.