Segmentation de données de type intervalle, diagramme et taxonomique

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. L'objectif de cet article consiste à étendre le critère de découpage binaire de Kolmogorov-Smirnov aux données de type intervalle, diagramme et taxonomique. Ce critère nécessite un ordre des valeurs prises par les variables explicatives. Nous utilisons différentes méthodes pour ordonner ce type de données. Nous présentons le format des questions binaires et la description des nœuds terminaux pour chacun de ces types de données. Nous étudions également la précision de ce critère et nous le comparons aux critères de l'entropie et de Gini.

1 Introduction

Dans le domaine de la discrimination par arbre de décision binaire, les variables explicatives sont souvent quantitatives ou qualitatives. Le critère de découpage binaire de Kolmogorov-Smirnov (noté KS dans la suite) a été introduit par (Friedman, 1977) pour une partition binaire à expliquer sur des variables continues. Ce critère a été exploré quelques années plutard par (Utgoff et Clouse, 1996) pour construire des arbres de décision binaires sur ce même type de données. Il a été étendu aux données qualitatives classiques par (Asseraf, 1998).

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 de découpage. Dans cet article, nous nous intéressons au critère de découpage binaire KS. C'est un critère présentant un bon pouvoir discriminant sur des données classiques. Nous étudions son extension dans le cas où les objets destinés à être classés par un arbre de décision sont décrits par des variables de type intervalle, diagramme et taxonomique.

Nous présentons tout d'abord différentes méthodes pour ordonner des données de type intervalle, diagramme et taxonomique. Ensuite, nous étudions l'extension du critère KS à ce type de données pour la construction d'arbres binaires de décision. Nous présentons également le format des questions binaires et la description des nœuds terminaux de l'arbre de décision pour chaque type de données. Enfin, nous comparons les critères KS, Gini et entropie sur des données de type intervalle et diagramme.