Arbres de décision sur des données de type intervalle : évaluation et comparaison

Chérif Mballo [*,**] & Edwin Diday [**]

[*] ESIEA Recherche, 38 Rue des Docteurs Calmette et Guérin 53000 Laval France
mballo@esiea-ouest.fr

[**] LISE-CEREMADE, Université Paris Dauphine, Place du Maréchal de Lattre de Tassigny, 75775 Paris cedex 16, France
diday@ceremade.dauphine.fr

Résumé. Le critère de découpage binaire de Kolmogorov-Smirnov nécessite un ordre total des valeurs prises par les variables explicatives. Nous pouvons ordonner des intervalles fermés bornés de nombres réels de différentes façons. Notre contribution dans cet article consiste à évaluer et à comparer des arbres de décision obtenus sur des données de type intervalle à l'aide du critère de découpage binaire de Kolmogorov-Smirnov étendu à ce type de données (Mballo et al. 2004). Pour ce faire, nous axons notre attention sur le taux d'erreur mesuré sur l'échantillon de test. Pour estimer ce paramètre, nous divisons aléatoirement chaque base de données en deux parties égales en terme d'effectif (à un objet près) pour construire deux arbres. Ces deux arbres sont d'abord testés par un même échantillon puis par deux échantillons différents.

1 Introduction

Dans le domaine de la discrimination par arbre de décision binaire, les variables explicatives sont souvent quantitatives ou qualitatives classiques. Le critère de découpage binaire de Kolmogorov-Smirnov a été introduit par (Friedman 1977 ; Utgoff et Clouse 1996) pour une partition binaire à expliquer avec des variables explicatives quantitatives classiques. Ce critère a été étendu aux variables explicatives qualitatives classiques par (Asseraf 1998). Cependant, depuis quelques années, 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 symboliques, notamment de type intervalle et histogramme (Périnel 1996 ; Yapo 2002). Ces auteurs utilisent les critères de découpage classiques (entropie, Gini, gain ratio, likelihood) pour construire l'arbre de décision. Nous privilégions ici la méthode basée sur le critère de découpage binaire de Kolmogorov-Smirnov. Ce critère est basé sur un ordre total des valeurs prises par les variables explicatives. Nous pouvons ordonner des intervalles fermés bornés de $\Re$ (ensemble des nombres réels) de différentes façons (Diday et al. 2003) et chacune des relations d'ordre proposées est totale sur l'ensemble des intervalles fermés bornés. Nous présentons ici une approche exploratoire de construction d'arbres de décision. Cette approche consiste à construire un arbre pour chaque ordre et à comparer ces arbres obtenus selon le taux d'erreur réel mesuré sur l'échantillon de test. Pour estimer ce paramètre, nous utilisons l'approche suivante : chaque base de données utilisée est divisée aléatoirement en deux parties pour construire deux arbres et ces arbres sont d'abord testés par un même échantillon puis par deux échantillons différents (section 5). Comme les