Classement d'objets incomplets dans un arbre de décision probabiliste

Lamis Hawarah, Ana Simonet , Michel Simonet*
*TIMC-IMAG
Institut d'Ingénierie et de l'information de Santé
Faculté de Médecine – IN3S
38700 LA TRONCHE
{Lamis.Hawarah, Ana.Simonet, Michel.Simonet}@imag.fr
http://www-timc.imag.fr

Résumé. Nous présentons une approche probabiliste pour déterminer les valeurs manquantes des objets incomplets pendant leur classement dans les arbres de décision. Cette approche est dérivée de la méthode d'apprentissage supervisé appelée Arbres d'Attributs Ordonnés proposée par Lobo et Numao en 2000, qui construit un arbre de décision pour chacun des attributs, selon un ordre croissant en fonction de l'information mutuelle entre chaque attribut et la classe. Notre approche étend la méthode de Lobo et Numao d'une part en prenant en compte les dépendances entre les attributs pour la construction des arbres d'attributs, et d'autre part en fournissant un résultat de classement d'un objet incomplet sous la forme d'une distribution de probabilités (au lieu de la classe la plus probable).

1 Introduction

Le problème des valeurs manquantes est un problème connu dans le domaine de la fouille de données, où, dans la base d'apprentissage, on rencontre des objets ayant des valeurs manquantes pour certains attributs. Nous étudions ce problème dans le cadre des arbres de décision. Un arbre de décision est construit à partir d'un ensemble d'apprentissage selon l'approche divide-and-conquer (Quinlan 1993). Une fois l'arbre construit, il est utilisé pour classer de nouveaux objets. Pour cela on parcourt l'arbre en commençant par la racine et en suivant les branches correspondant aux valeurs de l'objet, jusqu'à une feuille. La classe associée à cette feuille est la classe de cet objet. Les arbres de décision sont confrontés au problème des données manquantes, à la fois lors de leur construction et lors du classement d'objets. Lors de la construction, l'existence de valeurs manquantes pose problème pour le calcul de gain d'information, nécessaire au choix de l'attribut test, ainsi que pour la partition de l'ensemble d'apprentissage selon l'attribut test choisi. Le classement d'un objet avec des valeurs manquantes soulève également des problèmes lorsqu'un nœud correspondant à un attribut manquant est rencontré dans le parcours de l'arbre. Dans ce travail nous nous intéressons exclusivement au second problème, c'est-à-dire le classement d'objets incomplets.

Les méthodes qui traitent les valeurs manquantes dans les arbres de décision, remplacent un attribut manquant par une seule valeur, qui peut être la valeur la plus probable (Kononenko et al. 1984) ou la plus similaire (Breiman et al. 1984), etc. Ce type d'approche présente l'inconvénient d'oublier les autres valeurs possibles. Notre approche vise à une détermination probabiliste des valeurs manquantes, en prenant en compte les dépendances entre l'attribut manquant et les autres attributs de l'objet, ce qui permet d'utiliser le maximum de l'information contenue dans l'objet pour le calcul des valeurs manquantes. De plus, nous voulons un résultat sous la forme d'une distribution de probabilités plutôt que la valeur la plus probable, ce qui donne une information plus fine. Parce que les arbres de décision sont