Une approche probabiliste pour le classement d'objets incomplets dans un arbre de décision

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

Nous étudions le problème du classement d'objets incomplètement connus dans les arbres de décision. La question des valeurs manquantes dans les arbres de décision a surtout été considérée lors de la construction de l'arbre. Or, cette question est également très présente lors de son exploitation pour le classement d'objet. En effet, dans beaucoup de domaines, et en particulier en médecine, les objets à identifier ne sont en général que partiellement connus.

Notre approche est dérivée de la méthode d'apprentissage supervisé appelée Arbres d'Attributs Ordonnés (AAO), proposée par [Lobo et Numao, 2000]. Leur méthode utilise l'Information Mutuelle (IM), calculée entre chaque attribut et la classe. Un arbre de décision, appelé *arbre d'attribut*, est construit pour chaque attribut en commençant par l'attribut ayant l'IM minimale relativement à la classe. Le premier arbre construit est constitué d'un seul nœud, l'attribut, avec sa valeur la plus fréquente. Les arbres d'attributs sont calculés successivement, par ordre d'IM croissante avec la classe, chaque arbre n'utilisant que les attributs pour lesquels un arbre a déjà été construit. [Lobo et Numao, 2001] ont étudié les conditions que les données doivent satisfaire pour que cette méthode soit applicable.

Nous proposons deux aménagements à leur méthode : 1) pour chaque attribut, son arbre d'attribut est construit en utilisant les attributs avec lesquels il a une dépendance (IM non nulle), au lieu d'une construction sur le critère d'IM croissante relativement à la classe, mais sans prendre en compte les dépendances entre les attributs ; 2) on associe à un attribut dont la valeur est inconnue, non plus sa valeur la plus probable, mais l'ensemble de ses valeurs possibles, avec leur probabilité. En conséquence, le résultat du classement d'un objet incomplètement connu n'est plus une valeur unique de la classe, mais une distribution de probabilités de ses valeurs possibles. Pour cela, nous construisons des arbres d'attributs probabilistes (AAP), en utilisant les sous-ensembles de l'ensemble d'apprentissage initial constitué des attributs dépendants de l'attribut concerné. Les feuilles de cet arbre associent une probabilité à chaque valeur possible de l'attribut.

Les premiers résultats montrent une meilleure réponse des AAP dans les cas où les AAO n'étaient pas satisfaisants, et une étude plus complète est en cours. Le principal problème qui reste posé est l'existence possible de cycles dans les dépendances. Actuellement, en cas de cycle nous utilisons les AAO, enrichis d'information probabiliste. De plus, la prise en compte d'un seuil minimal d'IM pour les dépendances diminuera le risque de cycle.

Références

[Lobo et Numao, 2000] O.O. Lobo et M. Numao. Ordered estimation of missing values for propositional learning. Japanese Society for Artificial Intelligence, vol. 15, no.1, 2000.

[Lobo et Numao, 2001] O.O. Lobo et M. Numao. Suitable Domains for Using Ordered Attribute Trees to Impute Missing Values. IEICE TRANS. INF. & SYST., vol. E84-D, no.2, 2001.