L'arbre à nœuds probabilistes:

Une nouvelle approche à la construction d'arbres de prédictions.

Antonio Ciampi
Université McGill, Montréal, Québec, Canada
antonio.ciampi@mcgill.ca

Résumé : Nous montrons la connexion entre, d'une part, l'analyse de données symboliques, et, d'autre part, certains algorithmes d'apprentissage supervisé pour la prédiction d'une variable réponse qualitative ou quantitative. Dans le contexte des données symboliques, nous avions développé un algorithme de prédiction à arbre; cela nous avait permis de traiter des données imprécises et de construire des arbres de prédiction classiques à partir de ce type de données. Par la suite, nous avons repris le problème de la construction d'arbres pour des données précises (numérique), mais en permettant des nœuds probabilistes ou 'tendres', c'est-à-dire des nœuds correspondant à des décisions probabiliste du type : 'aller à gauche avec probabilité *p* et à droite avec probabilité 1-*p*'. Un tel arbre décrit la distribution prédictive conditionnelle de la variable réponse comme un mélange de distributions, tel que les coefficients des composantes du mélange dépendent des variables : ces coefficients sont en effet des produits de fonctions sigmoïdes de certaines variables de prédiction choisies par l'algorithme guidé par les données. Nous décrivons une approche EM pour l'estimation des paramètres du modèle correspondant. La méthode a été évaluée par des simulations et des analyses de données réelles. Nous discutons, pour conclure, les avantages et les limites de ce type d'arbre en comparaison avec les arbres conventionnels.

Mots clés : apprentissage supervisé, prédictions, arbre à nœuds probabilistes, données symboliques, données imprécises.

Abstract: We show the connection between symbolic data analysis (SDA) and certain algorithms of supervised learning for the prediction of a continuous or categorical outcome. In the context of SDA, we had previously developed a tree-growing algorithm which allowed us to handle imprecise data and to construct classical prediction trees from such data. Later we went back to tree-growing for classical (numerical) data, and proposed the notion of probabilistic or soft node, that is a node representing a decision of the type: 'go left with probability *p* and go right with probability 1-*p*'. Such a tree-shaped predictor describes the conditional predictive distribution of the outcome as a mixture of distributions, with mixing coefficients which are functions of certain predictor variables chosen by the algorithms guided by the data. We describe an EM approach to the estimation of the predictive model parameters. The method is evaluated by simulation and real data analyses. In conclusion, we discuss the advantages and the limitations of the tree with soft nodes in comparison with conventional prediction trees.

Keywords: supervised learning, prediction, trees with soft nodes, symbolic data analysis, imprecise data.