UNIVERSITÉ DE LIÈGE
Faculté des Sciences Appliquées
Département d'Électricité, Électronique et Informatique
Institut Montefiore

Contributions to Decision Tree Induction:

Bias/Variance Tradeoff
and
Time Series Classification

Thèse présentée par
Pierre Geurts
en vue de l'obtention du titre
de Docteur en Sciences Appliquées

Année Académique 2001–2002

Summary

Because of the rapid progress of computer and information technology, large amounts of data are nowadays available in almost any field of application. Automatic learning aims at producing synthetic high-level information, or models, from data. Learning algorithms are generally evaluated according to three different criteria: interpretability (how well the model helps to understand the data), predictive accuracy (how well the model can predict unseen situations), and computational efficiency (how fast the algorithm is and how well it scales to very large databases). Data mining refers to the application of automatic learning and visualization techniques in order to help a human expert to extract knowledge from large volumes of raw data. Unlike in other applications of automatic learning, in the context of data mining usually only a small fraction of the available data is actually relevant, and also, there is seldom reliable domain knowledge to help extracting relevant features or models. Actually, the precise objective of data mining is to help creating such knowledge from the available data.

Decision tree induction is a supervised learning algorithm, which focuses on the modeling of input/output relationships in the form of if-then rules. It has been claimed that among the various classes of learning methods, decision trees come closest to meeting the requirements for serving as an off-the-shelf procedure for data mining. Indeed, they are quite fast, produce relatively interpretable models, are immune to outliers, resistant to irrelevant variables, insensitive to variable rescaling, and can be easily extended to cope with a large variety of data types (numerical, symbolic, time series...). However, in terms of accuracy they seldom are competitive with other automatic learning algorithms, such as neural networks and nearest neighbors. It is commonly admitted that this suboptimality is mostly due to the excessive variance of this method. The variance of an automatic learning method measures the dependence of the models it produces on the random nature of the data set used (sampling and noise). In the context of decision trees, this variance affects all the parameters of the extracted rules and is detrimental in terms of both accuracy and interpretability. Furthermore, this variance tends to increase very strongly in applications such as time series classification where the input variables correspond to a large number of low level and, sometimes, very noisy features.

Within this overall context, this thesis explores two complementary issues: the reduction of variance of decision tree based models and the problem of learning interpretable and accurate classification rules from time series data. All our developments and analyses are extensively validated in an empirical fashion through the estimation of variance, bias, and average accuracy, on the basis of a diversified set of databases.

We start our investigations by an empirical study which shows quantitatively how important decision tree variance is, i.e. how strongly they depend on the random nature of the database used to infer them. These experiments confirm that the rules induced for a given problem are indeed highly variable from one learning sample to another, and thus that variance is detrimental not only for accuracy but also from the point of view of interpretability. With the goal of improving both interpretability and accuracy, we

then investigate three complementary variance reduction techniques. First, we propose several methods to stabilize the parameters chosen during tree induction. While these methods succeed in reducing the variability of the parameters, they produce only a slight improvement of accuracy. Next, we consider perturb and combine algorithms, which aggregate predictions of several models obtained by randomizing in some way the learning process. Among the existing such methods, the most popular ones, namely Bagging and Boosting, strongly improve tree accuracy but jeopardize interpretability and computational efficiency. Thus, inspired by the high variance of tree parameters, we propose a new algorithm based on extremely randomized trees (extra-trees). This method aggregates the predictions of many large trees, each one obtained by choosing most of its parameters fully at random. These extra-trees compare very favorably with both bagging and boosting, in terms of variance reduction, accuracy and computational efficiency. They are thus largely more accurate than standard trees, while remaining competitive with them in terms of computational efficiency. In order to also recover interpretability, we then propose a "dual" perturb and combine algorithm which randomizes input attributes at the prediction stage while using one single model. In combination with decision trees, this method actually yields soft decisions and eventually bridges the gap between single trees and the ensembles of trees produced by the perturb and combine methods. Of the first, it saves most of the interpretability, and with perturb and combine algorithms, it shares part of the accuracy improvement.

The last part of our work concerns a first attempt to design general algorithms solving the specific, but very multifaceted problem of time series classification. The most direct way to solve this problem is to apply existing learning algorithms on low-level variables which correspond to the values of a time series at several time points. Experiments with the tree-based algorithms studied in the first part of the thesis show that this approach has its limitations. We thus design a variance reduction technique specifically for this kind of data, which consists in aggregating the predictions given by a classification model based on small subsequences of the original time series. This method is very successful in terms of accuracy but does not provide fully interpretable models. We therefore propose a second method which extends decision tree tests by allowing them to detect local shift invariant properties, or patterns, in time series. This method, although limited in scope, offers the possibility to extract interpretable and accurate rules from time series.

Acknowledgements

(in french)

En premier lieu, je voudrais remercier Louis Wehenkel, Professeur à l'université de Liège et promoteur de cette thèse. Depuis l'encadrement de mon travail de fin d'études jusqu'à l'aboutissement de cette recherche quatre ans plus tard, il a toujours su me fournir le soutien scientifique et moral dont j'avais besoin. Je le remercie pour sa patience, sa disponibilité permanente, malgré un emploi du temps chargé, et la confiance qu'il a toujours su me témoigner même dans les moments où je doutais de tout.

Je remercie également le Professeur Mania Pavella d'avoir cru très tôt en moi, de m'avoir accueilli dans son service en tant qu'étudiant-moniteur et de m'avoir ainsi permis de découvrir les recherches qui y étaient menées en apprentissage automatique.

Je remercie sincèrement les Professeurs Pierre Wolper, Pascal Gribomont, Rodolphe Sépulchre, Patrick Gallinari, Gilbert Ritschard et Djamel Zighed d'avoir accepté de faire partie de mon jury de thèse. Je leur suis très reconnaissant de l'intérêt qu'ils portent à mon travail.

Mes remerciements vont aussi à toutes les personnes avec qui j'ai eu des discussions lors de différentes rencontres, conférences ou séminaires auxquels j'ai pu participer, et qui ont d'une manière ou d'une autre influencé l'orientation que j'ai donnée à ma recherche. Il m'est impossible de les citer tous ici. Je remercie aussi Luc Moreau et tous les membres du "Multimedia Research Group" de l'université de Southampton pour leur accueil lors d'un séjour de deux mois que j'ai accompli au sein de leur équipe en avril 1999. Je leur dois entre autres d'avoir osé écrire cette thèse en anglais.

Je suis très reconnaissant au Fonds National de la Recherche Scientifique de Belgique (F.N.R.S) pour la bourse d'aspirant dont j'ai bénéficié ces quatre dernières années ainsi que pour les crédits qui m'ont été octroyés pour des déplacements à diverses conférences.

Je remercie également tous mes collègues et amis du service de méthodes stochastiques, tant pour les échanges scientifiques que pour l'excellente ambiance de travail. Notre équipe ne tournerait certainement pas aussi efficacement sans l'aide et le dévouement de Marie-Berthe Lecomte et de Claude Diépart que je remercie. Je n'oublie pas non plus le groupe de la pause café du matin ou du repas de midi (que j'ai malheureusement dû négliger dans les derniers moments de la rédaction): Damien, Christophe en son temps, Jean-Marc, Raphaël, Cristina, Philippe, Ibtissam...

Enfin, et surtout, je voudrais remercier ma famille et mes amis qui m'ont toujours soutenu, sans trop bien savoir ce que je faisais et où tout cela allait me mener. Je ne remercierai jamais assez mes parents pour tout ce qu'ils ont fait pour moi. Ibtissam mérite certainement une pensée toute particulière, pour son aide, pour son soutien moral, pour sa patience et pour ce qu'elle est pour moi. Je lui dédie cette thèse même si je sais que cela ne remplacera jamais un voyage de noces.

Et bien sûr, je remercie aussi...tous ceux que j'ai oublié de remercier. Qu'ils m'en excusent.

Contents