Moyennage du prédicteur Bayesien Naïf Sélectif, évaluation sur un challenge international

Marc Boullé

France Telecom R&D,
2, avenue Pierre Marzin, 22300 Lannion
marc.boulle@francetelecom.com

Résumé : Le prédicteur Bayesien Naïf s'est avéré très compétitif sur de nombreux jeux de données réels. Ses performances bénéficient généralement d'une estimation précise des probabilités conditionnelles univariées et d'une sélection de variable efficace. Néanmoins, bien que la sélection de variables soit souhaitable, elle est sujette au sur-apprentissage. Dans ce papier, on introduit une nouvelle technique de régularisation permettant de choisir le sous-ensemble de variables le plus probable et on propose une nouvelle méthode de moyennage de modèle. Le schéma de pondération sur les modèles se traduit par un schéma de pondération sur les variables, ce qui finalement produit un prédicteur Bayesien Naïf avec une sélection "douce" de variables. La méthode a été évaluée à l'issue d'une participation à un challenge international et a montré des performances de premier plan.

Mots clés : Data Mining, Naïve Bayes, Sélection de variable, Regularisation, Combinaison de modèle.

1 Description de la méthode

Notre méthode est basée sur l'hypothèse Bayesienne naïve, d'indépendance des variables endogènes conditionnellement aux valeurs exogènes. Les variables endogènes sont discrétisées ou groupées selon leur type au moyen des méthodes MODL (Boullé, 2005; 2006a). Ces méthodes sont optimales au sens de Bayes, et fournissent des estimateurs univariés de densité conditionnelle fiables et précis. Ces estimateurs étant supposés fixés, l'apprentissage d'un prédicteur Bayésien Naïf Sélectif (Langley et Sage, 1994) se résume à la recherche d'un sous-ensemble de variables. Ce problème est formulé comme un problème de sélection de modèle, et résolu selon une approche Bayesienne. On propose une distribution a priori sur l'espace des sous-ensembles de variables, par un choix uniforme du nombre de variables, puis le nombre de variables étant défini, par un choix uniforme du sous-ensemble de variable dans un schéma avec remise. La vraisemblance conditionnelle est estimée au moyen du prédicteur Bayesien Naïf exploitant les variables sélectionnées. La recherche du modèle le plus probable est effectuée en alternant des passes de sélection avant et d'élimination arrière sur des listes de variable ordonnées aléatoirement. Un moyennage de modèle est effectué en s'inspirant d'une approche Bayesienne (Hoeting et al., 1999). La distribution a posteriori des modèles est lissée

par un logarithme, ce qui revient à utiliser les taux de compression informationnels des modèles pour le schéma de pondération. Le moyennage est calculé le long de la trajectoire d'optimisation des modèles, ce qui permet de garantir une complexité algorithmique efficace, super-linéaire en la taille des données (variables et instances).

La méthode complète est détaillée dans (Boullé, 2006b).

2 Evaluation sur un challenge international

Le Performance Prediction Challenge (Guyon, 2006) est une compétition portant sur 5 jeux de données, destinée à stimuler la recherche et à révéler l'état de l'art en matière de sélection de modèle. Le challenge s'est étalé du 30/09/2005 au 01/03/2006 et a attiré 145 participants pour un total de plus de 4000 soumissions. Seuls les 28 participants ayant effectué une soumission complète sur les 5 jeux de données ont été évalués au terme du challenge, sur le critère du taux d'erreur équilibré (BER).

Notre méthode se place 7^{ième} globalement parmi les participants. Sur 2 des 5 jeux de données (ADA et SYLVA), notre meilleure soumission se place 1^{ière}.

Table 1. Evaluation de la méthode sur les jeux de données du challenge

Dataset	Notre meilleure soumission				La soumission gagnante du challenge			
	Test AUC	Test BER	Ber Guess	Test score	Test AUC	Test BER	Ber Guess	Test score
ADA	0.9149	0.1723	0.1650	0.1793	0.9149	0.1723	0.1650	0.1793
GINA	0.9772	0.0733	0.0770	0.0767	0.9712	0.0288	0.0305	0.0302
HIVA	0.7542	0.3080	0.3170	0.3146	0.7671	0.2757	0.2692	0.2797
NOVA	0.9736	0.0776	0.0860	0.0858	0.9914	0.0445	0.0436	0.0448
SYLVA	0.9991	0.0061	0.0060	0.0062	0.9991	0.0061	0.0060	0.0062
Overall	0.9242	0.1307	0.1306	0.1399	0.8910	0.1090	0.1040	0.1165

Références

BOULLÉ M. (2005). A Bayes Optimal Approach for Partitioning the Values of Categorical Attributes. *Journal of Machine Learning Research* 6. p. 1431-1452.

BOULLÉ M. (2006a). MODL: a Bayes Optimal Discretization Method for Continuous Attributes. *Machine Learning*. to be published.

BOULLÉ M. (2006b). Regularization and Averaging of the Selective Naïve Bayes classifier. *International Joint Conference on Neural Networks*. to be published.

GUYON I. (2006). Model Selection Workshop and Performance Prediction Challenge, IEEE World Congress on Computational Intelligence, Vancouver 2006, http://clopinet.com/isabelle/Projects/modelselect/#challenge.

HOETING J.A., MADIGAN D., RAFTERY A.E. and VOLINSKY C.T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, 14(4). P. 382-417.

LANGLEY P. and SAGE S. (1994). Induction of Selective Bayesian Classifiers. *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann. p. 399-406.