Classifieur naïf de Bayes pondéré pour flux de données

Christophe Salperwyck*, Vincent Lemaire**, Carine Hue**

*Powerspace, 13 rue Turbigo, 75002 Paris
** Orange Labs, 2 avenue Pierre Marzin, 22300 Lannion

Résumé. Un classifieur naïf de Bayes est un classifieur probabiliste basé sur l'application du théorème de Bayes avec l'hypothèse naïve, c'est-à-dire que les variables explicatives (X_i) sont supposées indépendantes conditionnellement à la variable cible (C). Malgré cette hypothèse forte, ce classifieur s'est avéré très efficace sur de nombreuses applications réelles et est souvent utilisé sur les flux de données pour la classification supervisée. Le classifieur naïf de Bayes nécessite simplement en entrée l'estimation des probabilités conditionnelles par variable $P(X_i|C)$ et les probabilités a priori $P(C)$. Pour une utilisation sur les flux de données, cette estimation peut être fournie à l'aide d'un « résumé supervisé en-ligne de quantiles ». L'état de l'art montre que le classifieur naïf de Bayes peut être amélioré en utilisant une méthode de sélection ou de pondération des variables explicatives. La plupart de ces méthodes ne peuvent fonctionner que hors-ligne car elles nécessitent de stocker toutes les données en mémoire et/ou de lire plus d'une fois chaque exemple. Par conséquent, elles ne peuvent être utilisées sur les flux de données. Cet article présente une nouvelle méthode basée sur un modèle graphique qui calcule les poids des variables d'entrée en utilisant une estimation stochastique. La méthode est incrémentale et produit un classifieur Naïf de Bayes Pondéré pour flux de données. Cette méthode est comparée au classique classifieur naïf de Bayes sur les données utilisées lors du challenge « Large Scale Learning ».

1 Introduction

Pour les données hors-ligne, des méthodes d'extractions de connaissances performantes et éprouvées depuis plusieurs années existent. Différents types de classifieurs ont été proposés : plus proches voisins, bayésien naïf, SVM, arbre de décision, système à base de règles... Mais avec l'apparition de nouvelles applications comme les réseaux sociaux, la publicité en-ligne, les données du Web... la quantité de données et leurs disponibilités ont changé. Les données auparavant facilement disponibles et pouvant tenir en mémoire (données hors-ligne) sont devenus massives et visibles une seule fois (flux de données). La plupart des classifieurs, prévus pour fonctionner hors-ligne, ne peuvent généralement pas s'appliquer directement sur un flux de données.

Depuis les années 2000, l'extraction de connaissances sur flux de données est devenue un sujet de recherche à part entière. De nombreux travaux traitant cette nouvelle problématique ont été proposés (Salperwyck et Lemaire, 2011; Gama, 2010). Parmi les solutions aux problèmes