Elaboration d'une représentation basée sur un classifieur et son utilisation dans un déploiement basé sur un k-ppv

Vincent Lemaire, Marc Boullé, Pascal Gouzien
Orange Labs 2 avenue Pierre Marzin 22300 Lannion
http://perso.rd.francetelecom.fr/lemaire

Résumé : Dans le cadre industriel de cet article, on cherche à utiliser l'algorithme des k-ppv pour déployer rapidement un classifieur existant par ailleurs. Le classifieur existant est utilisé comme source d'information pour construire une métrique qui lui est adaptée. Puis un k-ppv est construit de manière à accélérer notablement le déploiement en comparaison d'un déploiement direct.

Mots-clés : Projection, Classifieur, K-ppv, Représentation

1 Introduction

1.1 Problématique industrielle

La manière la plus classique de construire et exploiter de l'information dans un système de gestion de relation client est de produire des scores. Un score est une valeur en sortie d'un modèle prédictif, qui exploite un grand nombre de variables explicatives issues du système d'information client afin de prédire un comportement client, comme par exemple le churn ou l'appétence aux nouveaux produits. Les scores sont alors utilisés par les services marketing pour personnaliser la relation client. Pour faciliter le déploiement d'un classifieur (M), (Féraud *et al.* (2008)) ont proposé d'extraire de la base de données, une table de parangons. La table des parangons contient les individus représentatifs des variables explicatives utilisées par le modèle. Les parangons sont reliés par un index à toute la population : chaque individu est associé à son parangon le plus proche.

Lors d'un déploiement classique, sans la table d'index, le processus de déploiement inclut 2 étapes principales : (i) la base de données $T1$ ($K1$ instances représentées par J variables explicatives) est extraite du système d'information (SI) et (ii) le classifier (M) est appliqué sur l'ensemble des instances de cette base de données. Dans la plateforme industrielle, décrite dans (Féraud *et al.* (2008)), les données présentent dans le SI sont structurées et stockées dans une base de données relationnelle. L'extraction consiste en (i) construction des variables explicatives à l'aide de jointures entre les différentes tables du SI relationnel (ii) élaboration d'une table à plat : instances x variables explicatives. Le coût de déploiement est l'addition du coût d'extraction (C_{e1}) et du coût de

classification (C_{c1}). Si le problème de classification est supposé stationnaire, le classifieur n'a pas besoin d'être réappris et peut être à nouveau déployé à l'aide de ces 2 étapes, seul le contenu de $T1$ à changé.

Pour faire baisser ce coût de 'déploiement direct' lors des déploiements ultérieurs, 2 tables sont élaborées et stockées lors du premier déploiement : (i) une table de parangons, $T2$ ($K2$ instances représentées par J variables explicatives) à l'aide d'un tirage aléatoire dans $T1$ et (ii) une table d'index qui contient pour chaque instance de $T1$ ses plus proches voisins dans $T2$. Les constructions de ce deux tables ont respectivement comme coût C_{rs} et C_{id}.

Lors du deuxième déploiement, le classifieur est déployé sur $T2$ pour obtenir la classification des parangons (avec un coût C_{c2}). La classification de toutes les instances de T1 est obtenu par une simple jointure entre la table des parangons et la table d'index (avec un coût C_{jo}) : chaque instance de $T1$ reçoit le score de ses plus proches voisins de $T2$. Cette approche est particulièrement efficace lors du déploiement récurrent du classifieur à déployer. Le gain est 2 ordre : (1) dans l'étape d'extraction ($K2 \ll K1$) et (2) dans l'étape de déploiement car le temps pour calculer la jointure entre la table d'index et la table des parangons est plus petit ($(C_{c2} + C_{jo}) \ll C_{c1}$) que l'application du classifieur sur toute la population ($T1$).

Dans ce cadre industriel très particulier, le point clef réside dans la construction de la table d'index. Le problème est alors la construction d'un déploiement basé sur un k-ppv (entre $T1$ et $T2$) avec une perte minimum de performance en comparaison d'un déploiement direct (du classifieur à déployer). Il est important de noter ici que dans la suite de l'article on ne va pas chercher à construire un nouveau classifieur basé sur un k-ppv. Le but va être de construire une métrique pour qu'un déploiement basé sur la table d'index, et donc sur un k-ppv, réalise au mieux le déploiement du classifieur initial (M).

2 Description et positionnement

Nous proposons dans cet article de projeter les instances dans l'espace des importances des variables explicatives en entrée du classifieur à déployer puis de construire un k-ppv basé sur cette nouvelle représentation. Etant donné :

- une table de modélisation T contenant K instances et J variables explicatives ;
- un problème de classification à C classes ;
- un classifieur probabiliste M entraîné sur la table de modélisation de manière à réaliser une classification ;
- une méthode de calcul, G, permettant, connaissant M, de calculer l'importance d'une variable en entrée du classifieur instance par instance ;
- une instance x_k représentée sous la forme d'un vecteur à J dimension : $x_k = (V_1 = x_{1k}, ..., V_J = x_{Jk})$

Après apprentissage M classe les instances contenues dans T (ou dans une autre table de déploiement) tel qu'a toute instance (k) on fait correspondre une classe d'appartenance. Dès lors muni de M, T et G on peut projeter les instances contenues dans T, dans l'espace des importances du classifieur en utilisant M et G. On nomme I cette réprésentation tel que $I_j^z(x_k)$ est l'importance pour l'instance x_k de la variable j étant donné la classe z.

Quelque soit l'indicateur d'importance utilisé on a alors un exemple x_k qui est représenté par le vecteur des importances de ses variables et ce pour toutes les classes. Le nombre de dimensions initiales de x_k est donc multiplié par le nombre de classes, C, du problème de classification à résoudre. La distance séparant deux instances correspond à la norme L1 pour tous les indicateurs présentés ci-dessus tel que .

$$\text{Dist}(x_k, x_k^{'}) = \sum_{j=1}^{J} \sum_{z=1}^{C} \left| I_j^z(x_k) - I_j^z(x_k^{'}) \right| \quad (1)$$

3 Impact de la représentation sur un k-ppv

3.1 Le cas du classifieur naïf Bayes - Représentations étudiées

Représentation générique : On utilise ici un des indicateurs d'importance décrit dans (Robnik-Sikonja & Kononenko (2008)) et qui fait référence en terme d'indicateurs d'importance pour un classifieur naïf de Bayes :

1 "Weight of Evidence (WOE)" : $I_j^z = log_2\left(odds(C_z|x_k)\right) - log_2\left(odds(C_z|x_k \backslash V_j\right)$;

Représentation supervisée sans modèle : On introduit 3 indicateurs supervisés univarié :

2 "Modality probability (MOP)" : $I_j^z = P(V_j = x_{jk}|C_z)$;

3 "Log Modality probability (LMOP) : $I_j^z = log(P(V_j = x_{jk}|C_z))$;

4 "Minimum of variable probabilities difference (VPD)" : $I_j^z = P(V_j = x_{jk}|C_z) - \max_{q \neq z} P(V_j = x_{jk}|C_q)$;

Représentation dédiée au modèle : Dans le cas du classifieur naïf de Bayes moyenné il est possible d'écrire une représentation exacte connaissant les paramètres du modèle. La représentation (MOLD) est alors définie par :

5 $I_j^z = log_2(P(V_j = x_{jk}|C_z))W_j$, où W_j est le poids de la variable j selon le classifieur naïf de Bayes moyenné

Le but des expérimentations sera de montrer que les représentations génériques permettent d'obtenir des résultats similaires à ceux obtenus avec la représentation dédiée.

3.2 Experimentations

On utilise les données du challenge KDD Cup 2009. Le but de challenge était de prédire la propension des clients à changer d'opérateur (churn) ou à acheter des services additionnels destinés à rentabiliser les ventes (up-selling). Dans la version 'small' du challenge, chaque problème de classification était constitué d'une base de données (anonymisée) contenant 100000 instances (clients) chacun décrit au moyen de 230 variables générées par un expert. La description complète du protocole du challenge ainsi que l'analyse des résultats peuvent être trouvées dans (Guyon *et al.* (2009)). On n'utilise ici que les données d'apprentissage challenge soit 50000 instances. Les pourcentages d'exemples positifs étaient respectivement pour les problèmes de churn et d'up-selling : 1.8% (890/50000) et 7.4% (3682/50000). Un 5-fold cross validation a été réalisé.

Le classifieur à déployer (M) est un classifieur naif de Bayes selectif (SNB) obtenu à l'aide du logiciel Khiops (`www.khiops.com`). Les probabilités conditionnelles ($P(V_j = x_{jk}|C_z)$), des variables continues et catégorielles, utilisées dans les 5 représentations de la section 3.1 ont été estimées à l'aide de (Boullé (2007)).

Les Figures 1 et 2 présentent respectivement les résultats obtenus en terme d'AUC moyen en test pour les problèmes de churn et d'upselling. Les représentations supervisées "sans modèle" (MOP, LMOP et VPD) fournissent des résultats inférieurs aux représentations supervisées issues du calcul de l'importance des variables explicatives présentes en entrée du classifieur à deployer (M). Ces figures montrent aussi que les représentations génériques (ici WOE), basées sur une analyse de sensibilité, atteignent les performances de la représentation dédiée (MODL) et sont proches du déploiement direct (SNB).

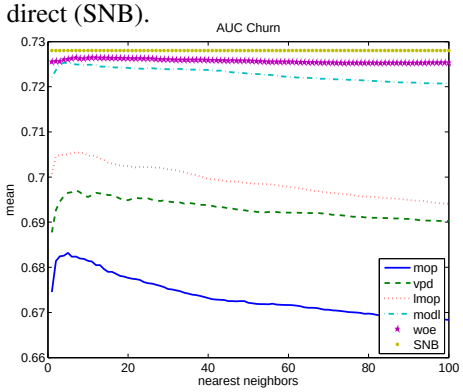

FIG. 1 – AUC Test Churn

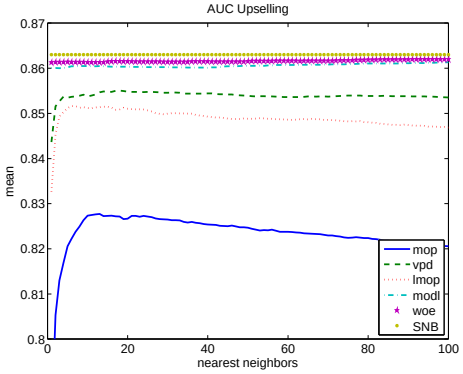

FIG. 2 – AUC Test Upselling

4 Conclusion

On a présenté dans cet article une méthode permettant de créer une représentation des données qui permet à un k-ppv d'atteindre les performances d'un classifieur initialement construit. Il a été montré que la méthode proposée est doublement générique : (1) les instances peuvent être projetées à l'aide d'une mesure de sensibilité non inhérente au classifieur préalablement construit (2) le classifieur préalablement construit peut être quelconque pour peu qu'il soit probabiliste. Il a été montré que sur 2 problèmes de marketing il est possible de créer une représentation efficace en termes de plus proches voisins. La performance obtenue égale celle du classifieur préalablement construit. Le problème industriel présenté en introduction, calculer la table d'index, reçoit une solution extrêmement efficace.

Références

BOULLÉ M. (2007). *Recherche d'une représentation des données efficace pour la fouille des grandes bases de données*. PhD thesis, Ecole Nationale Supérieure des Télécommunications.

FÉRAUD R., BOULLÉ M., CLÉROT F. & FESSANT F. (2008). Vers l'exploitation de grandes masses de données. In *Extraction et Gestion des Connaissances (EGC)*, p. 241–252.

GUYON I., LEMAIRE V., BOULLÉ M., DROR G. & VOGEL D. (2009). Analysis of the kdd cup 2009 : Fast scoring on a large orange customer database. *JMLR Workshop and Conference Proceedings*, **7**, 1–22.

ROBNIK-SIKONJA M. & KONONENKO I. (2008). Explaining classifications for individual instances. *IEEE TKDE*, **20**(5), 589–600.