Exploration des corrélations dans un classifieur Application au placement d'offres commerciales

Vincent Lemaire*, Carine Hue**

* Orange Labs, 2 avenue Pierre Marzin, 22300 Lannion
vincent.lemaire@orange-ftgroup.com,
http://perso.rd.francetelecom.fr/lemaire
**GFI Informatique, 11 rue Louis de Broglie, 22300 Lannion
chue@gfi.fr

Résumé. Cet article présente une nouvelle méthode permettant d'explorer les probabilités délivrées par un modèle prédictif de classification. L'augmentation de la probabilité d'occurrence de l'une des classes du problème étudié est analysée en fonction des variables explicatives prises isolément. La méthode proposée est posée et illustrée dans un cadre général, puis explicitement dédiée au classifieur Bayesien naïf. Son illustration sur les données du challenge PAKDD 2007 montre que ce type d'exploration permet de créer des indicateurs performants d'aide à la vente.

1 Introduction

Etant donné une base de données, une question classique est de chercher à relier un phénomène dit "à expliquer" à un ou plusieurs phénomènes explicatifs. L'extraction de connaissances passe alors couramment par l'élaboration d'un modèle qui explicite cette relation. Pour chaque individu de la base, un modèle probabiliste permet, étant données les valeurs de l'individu pour chaque variable explicative, d'estimer les probabilités d'occurrence de chaque classe cible ainsi que la classe cible prédite. Ces probabilités ou scores sont réinjectés dans le système d'information pour par exemple personnaliser la relation clients : le choix des offres, de l'interface des services, du canal de communication, du canal de distribution... Néanmoins la connaissance extraite sur un phénomène par le score n'est pas toujours exploitable directement. Par exemple, si un modèle prédit pour un client son potentiel à adhérer ou non à une offre, autrement dit son appétence à cette offre, il ne dit rien sur l'action ou les actions à entreprendre pour rendre plus probable son adhésion. Il semble ainsi nécessaire de posséder une méthodologie qui, pour chaque client, (i) permettra d'identifier l'importance des variables explicatives (ii) permettra d'identifier le placement des valeurs de ces variables explicatives et (iii) proposera d'entreprendre une action pour augmenter son appétence à l'offre.

Nous proposons de traiter ce troisième point en explorant la relation existante, au sens du classifieur, entre les variables explicatives prises indépendamment et la variable cible. Cette exploration, à réaliser pour chacun des clients, produit une connaissance qui sera ensuite exploitée dans un processus de "Customer Relationship Management" (CRM) pour, par exemple, fournir une information personnalisée dans l'argumentaire des téléopérateurs lors d'une campagne de promotion.

2 Positionnement

La mesure de l'importance d'une variable permet de sélectionner un sous-ensemble de variables pertinentes pour un problème donné. Cette sélection de variables (Guyon, 2005) permet d'augmenter la robustesse des modèles et de faciliter l'interprétation d'un modèle. L'importance de la valeur d'une variable consiste à analyser l'importance de la valeur prise par l'instance pour cette variable (Lemaire et al., 2008) (Robnik-Sikonja et Kononenko, 2008).

Cet article propose de compléter les deux aspects présentés ci-dessus. On se restreint dans cet article aux problèmes de classification, quel que soit le nombre de classes. On propose de s'intéresser à l'individualisation du lien de corrélation qui existe entre un exemple présenté à l'entrée d'un modèle et la probabilité d'apparition d'une classe en sortie du modèle. Pour un individu donné, la valeur d'une variable d'entrée prise isolément peut "tirer vers le haut" (valeur élevée) ou "tirer vers le bas" (valeur faible) la sortie du modèle. L'idée retenue est d'analyser la relation entre les valeurs d'une variable explicative et la probabilité d'apparition d'une des classes de la variable cible. En explorant les différentes valeurs prises par la variable d'entrée étudiée on cherchera par exemple à rendre plus appétant un client (augmenter la valeur de la sortie du modèle).

3 Exploration des valeurs prédictives

3.1 Le cas général

Soit C_z la classe cible d'intérêt parmi les T classes cible, par exemple l'appétence d'un client x à un produit ou son attrition (autrement dit son départ). Soit f_z la fonction qui modélise la probabilité d'occurrence de cette classe cible $f_z(X = x) = P(C_z|X = x)$ étant donné l'égalité du vecteur X des J variables explicatives à un vecteur donné x de J valeurs. La méthode proposée ici recherche à augmenter la valeur de $P(C_z|X = x_k)$ pour chacun des K exemples de l'échantillon.

Exploration : Pour l'exemple x_k, $P(C_z|x_k)$, est la valeur "naturelle" de la sortie du modèle. Nous proposons de calculer l'effet de la modification des valeurs de l'exemple x_k sur la sortie du modèle pour cet exemple. En pratique, on propose d'explorer les valeurs indépendamment pour chaque variable explicative. On note $P_j(C_z|x_k, b)$ la sortie du modèle f_z étant donné l'exemple x_k mais pour lequel on a remplacé la valeur de sa j^{ieme} composante et elle seule par une valeur b. Par exemple, si l'on modifie la valeur de la 3ème variable explicative parmi $J = 5 : P_3(C_z|x_k, b) = f_z(x_k^1, x_k^2, b, x_k^4, x_k^5)$). En parcourant l'ensemble des variables et pour chacune d'elles l'ensemble de ses valeurs on explore ce qu'aurait pu être la valeur de la sortie du modèle pour l'instance x_k en modifiant la valeur d'une des variables explicatives.

Quelles valeurs (b) essayer ? : L'intérêt de prendre en compte la distribution empirique des données a été montré expérimentalement (Lemaire et al., 2008). L'ensemble des valeurs utilisées pour les J variables explicatives seront donc les valeurs des K exemples disponibles dans la base de données qui sert à apprendre le classifieur : l'ensemble d'apprentissage. On pourra réduire ce nombre de manipulations aux ensembles de valeurs distinctes. On notera N_j le nombre de valeurs de la variable X_j.

Ordonnancement des explorations : On explore les variables explicatives une à une en parcourant l'ensemble des valeurs prises sur l'ensemble d'apprentissage. Les explorations étant

Pour l'exemple (le client) x_k **faire**
  w=0 ;
  Pour toutes les variables explicatives X_j **de** $j = 1$ **à** $j = J$ **faire**
    Pour toutes les n valeurs différentes (v_{jn}) de la variable X_j **de** $n = 1$ **à** $n = N_j$ **faire**
      Si $P_j(C_z|x_k, b = v_{jn}) > P(C_z|x_k)$ **Alors**
        $Ca[w] = v_{jn}$;
        $PCa[w] = P_j(C_z|x_k, v_{jn})$
        $XCa[w] = j$
      Sinon
        $Ca[w] = 0.0$;
        $PCa[w] = 0.0$;
        $XCa[w] = j$
      Fin Si
      w=w+1 ;
    Fin Pour
  Fin Pour
  Tri décroissant, selon les valeurs de $PCa[w]$, de $PCa[w]$, $Ca[w]$, $XCa[w]$.
Fin Pour

Algorithme 1: Exploration et ordonnancement des améliorations du score

indépendantes les unes des autres, l'ordre des explorations est quelconque. Lorsque la modification de la valeur d'une variable conduit à l'amélioration de la probabilité prédite on conserve (i) la valeur qui conduit à cette amélioration (Ca) (ii) la probabilité améliorée associée (PCa) et (iii) la variable associée à l'amélioration (XCa). On ordonne ensuite ces triplets selon l'amélioration obtenue sur la probabilité prédite. Notons qu'il est possible de ne pas parvenir à améliorer $P(C_z|x_k)$ et que dans ce cas les tables CA et PCa ne contiennent que des valeurs nulles.

Cas des changements de classe : Lors de l'utilisation de l'algorithme 1 on peut être amené à observer des changements de classes prédites. En effet il est d'usage d'utiliser la formulation suivante pour désigner la classe prédite pour l'exemple x_k : $\text{argmax}_z\,[P(C_z|x_k)]$. L'utilisation de l'algorithme 1 sur un exemple x_k appartenant à la classe $t(t \neq z)$ peut permettre d'obtenir $P(C_z|x_k, b) > P(C_t|x_k)$ et dans ce cas la cause correspondante (Ca) représente une cause riche d'information et exploitable. Par exemple dans le cas d'un problème d'attrition on cherchera à augmenter la probabilité qu'un individu reste client : les clients classés comme "churneur" par le classifieur et qui deviennent non-churneur via l'algorithme 1 seraient à contacter en priorité par le service commercial. On sait que ces clients risquent de partir mais on connaît une cause, valeur d'action, à réaliser pour qu'ils restent fidèles.

3.2 Déclinaison dans le cas d'un classifieur Bayesien

Un classifieur Bayesien naïf suppose que toutes les variables explicatives sont indépendantes sachant la classe cible. Cette hypothèse réduit drastiquement les calculs nécessaires. En utilisant le théorème de Bayes, l'expression de l'estimateur obtenu pour la probabilité conditionnelle d'une classe C_z est :

$$P(C_z|x_k) = \frac{P(C_z)\prod_{j=1}^{J} P(X_j = v_{jk}|C_z)}{\sum_{t=1}^{T}\left[P(C_t)\prod_{j=1}^{J} P(X_j = v_{jk}|C_t)\right]} \quad (1)$$

La classe d'appartenance est celle qui maximise les probabilité conditionnelles. En dépit de cette hypothèse d'indépendance, un tel classifieur présente des résultats satisfaisants (Hand et Yu, 2001). De plus, son expression permet bien d'explorer les valeurs des variables une à une indépendamment. Les probabilités $P(X_j = v_{jk}|C_z)$ $(\forall j, k, z)$ sont estimées (i) pour les variables continues par comptage après discrétisation et (ii) pour les variables modales par comptage après groupement de modalités (Boullé, 2008). Le dénominateur de l'équation 1 permet d'avoir $\sum_z P(C_z|x_k) = 1$.

Pour mesurer la fiabilité des informations produites par notre approche, nous l'avons testé sur des campagnes marketing de France Télécom[1] avec ou sans notre technologie. Nous avons été alimentés par la plateforme PAC (Féraud et al., 2008) avec différents jeux de données provenant des applications décisionnelles du groupe France Télécom. Nous avons consolidé des informations de plus de 1 000 000 de clients du groupe représentés chacun par plusieurs centaines variables explicatives. Ces tests nous ont amenés aux détails d'implémentation énumérés dans le document présent en ligne ici : `http://perso.rd.francetelecom.fr/lemaire/understanding/addon-khiops.pdf`, notamment quand au calcul de $P(C_z|x_k)$, et $P_j(C_z|x_k, b)$ nécessaire à l'utilisation de l'algorithme 1. Cette implémentation est alors "temps réel" sur l'écran d'un téléopérateur qui interrogerait l'application afin de connaître les actions à proposer à un client de manière à ce qu'il soit plus appétant par la suite à une offre commerciale.

4 Expérimentation - Application à la vente

4.1 Les données et conditions expérimentales

On utilise ici les données du challenge[2] qui a eu lieu lors de la conférence PAKDD en 2007. La société qui a fourni les données a actuellement une clientèle qui possède une carte de crédit aussi bien qu'une clientèle qui possède un prêt immobilier. Chacun de ces produits est sur le marché depuis de nombreuse d'années, mais, pour des raisons inconnues, le chevauchement entre ces deux clientèles est actuellement très petit. La société voudrait accroître ce chevauchement mais la petite taille du chevauchement présente un challenge lorsqu'on essaye d'élaborer un modèle de prédiction efficace pour prédire l'appétence d'un client (possesseur d'un produit) à l'autre produit.

Un ensemble de données de modélisation de 40700 clients avec 40 variables de modélisation (à partir du point d'application pour la carte de crédit de la société), plus une variable cible, a été fourni aux participants. Il s'agit d'un échantillon de clients qui ont ouvert une nouvelle carte de crédit avec la société dans une période spécifique de 2 ans et qui n'avaient pas de prêt immobilier existant avec la société. La variable catégorielle cible vaut 1 si le client a alors ouvert un prêt immobilier avec la société dans les 12 mois qui suivent l'ouverture de la carte de crédit (700 échantillons) et vaut 0 sinon (40000 échantillons). Les résultats des participants étaient comparés sur un ensemble de données de prédiction (sans la variable cible) de 8000 échantillons.

[1] Résultats non autorisés à la publication à ce jour.

[2] `http://lamda.nju.edu.cn/conf/pakdd07/dmc07/` : Les données ne sont plus en ligne mais les résultats et leur analyse y restent présentés. Nous remercions Mingjun Wei (participant référencé P049) pour nous avoir transmis les données (version 3).

Le challenge étant terminé il ne nous a pas été possible d'évaluation notre classifieur sur cet échantillon de prédiction. Nous avons alors élaboré un classifieur en utilisant les 40700 échantillons dans une procédure de validation croisée à 5 blocs. Nos résultats en classification ont été les suivants comparable à ceux obtenus par le tiercé gagnant ce challenge. Par exemple le meilleur résultat que nous avons obtenu sur l'un des 5 blocs est : AUC (Fawcett, 2003) Train = 68.82, AUC Test = 70.11.

4.2 Exploration

On choisit d'utiliser le meilleur classifieur obtenu en test au cours de la section précédente. Le modèle utilisé, le classifieur Bayesien naïf (Boullé, 2008), n'a sélectionné que 8 variables sur 40 : RENT_BUY_CODE, PRE_RES_MTHS, CURR_RES_MTHS, ENQ_L6M_GR3, B_ENQ_L3M, B_ENQ_L12M_GR3, B_ENQ_L12M_GR2, AGE_AT_APPLICATION.

Parmi ces variables, on exclut de l'exploration celles pour lesquelles il n'est pas envisageable de modifier leurs valeurs (telles que l'age, le sexe par exemple). On conserve les varaibles dites levier, c'est à dire celles sur lesquelles on pense pouvoir agir ou celles qui peuvent évoluer dans la vie du client (et sont donc "observables"). A titre d'exemple la variable "Residential_status_code" est non modifiable par une offre quelconque mais observable : si le client passe par exemple du groupe de valeur [O,P] ('O' Owner, 'P' Parents) à [M,R,B,X] ('M' Mortgage, 'R' Rent, 'B' Board, 'X' Other). Parmi les 8 variables sélectionnées par le modèle les 2è et 8è variables sont alors retirées de la liste des variables levier.

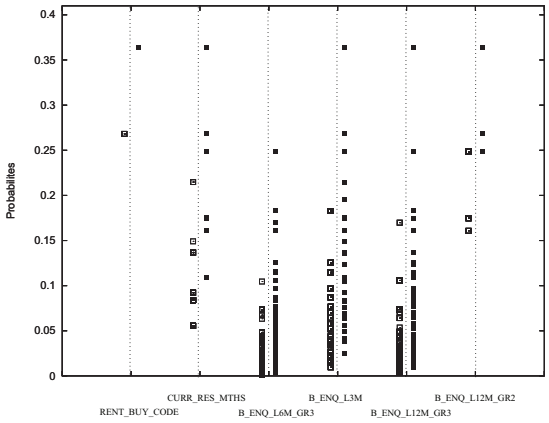

FIG. 1 – *Résultats sur les probabilités : à gauche de chaque axe pointillé vertical on trouve la distribution des $P(C_z|x_k)$ (□) et à droite du même axe la distribution des $P_j(C_z|x_k, b)$ (■). Les 6 axes pointillés posés verticalement représentent les 6 variables explicatives comme indiqué en abscisse. Les valeurs des probabilités sont indiquées selon l'axe des ordonnées. Pour chaque client seule la meilleure $P_j(C_z|x_k, b)$ ($PCa[0]$ dans l'algorithme 1) est tracée.*

La valeur "yes" de la variable cible est choisie comme classe d'intérêt (C_z). Il est à noter que cette classe est très faiblement représenté (1.75%). Le taux de classification présenté cidessus ou sur le site internet du challenge masque le fait que peu (ou pas selon nos découpages

en blocs) de clients sont classés “yes” par le classifieur. La manipulation des variables retenues ne permet pas de faire changer de classe les individus de notre ensemble de test. Néanmoins la figure 1 présente l'accroissement de la probabilité de vente croisée de l'entreprise en question.

5 Conclusion

On a présenté dans cet article une méthode permettant d'influer sur les probabilités délivrées par un modèle prédictif de classification en explorant les valeurs des variables explicatives pour chaque exemple. Elle a été illustré sur le challenge PAKDD 2007. Il a été montré que sur ce problème difficile de vente croisée il est possible de créer des indicateurs performants qui devraient permettre un accroissement des ventes. Cette méthode, simple mais très performante, est à ce jour intégrée dans un complément du logiciel Khiops^{TM}. Le manuel d'utilisation présentant l'interface peut être trouvé ici : `http://perso.rd.francetelecom.fr/lemaire/understanding/addon-khiops.pdf`. Cet outil serait utile aux entreprises et/ou organisations et/ou chercheur en apprentissage qui désirent comprendre les résultats de leur classification soit par une interprétation soit à l'aide de l'exploration de valeurs prédictives.

Références

Boullé, M. (2008). Khiops : outil de préparation et modélisation des données pour la fouille des grandes bases de données. In *Extraction et gestion des connaissances (EGC'2008)*, pp. 229–230.

Fawcett, T. (2003). Roc graphs: Notes and practical considerations for data mining researchers. Technical Report HPL-2003-4, HP Labs, 2003.

Féraud, R., M. Boullé, F. Clérot, et F. Fessant (2008). Vers l'exploitation de grandes masses de données. In *Extraction et Gestion des Connaissances (EGC)*, pp. 241–252.

Guyon, I. (2005). *Feature extraction, foundations and applications*, Chapter An Input Variable Importance Definition based on Empirical Data Probability, pp. 1–13. Elsevier.

Hand, D. et K. Yu (2001). Idiot's Bayes - not so stupid after all? *International Statistical Review 69*(3), 385–399.

Lemaire, V., R. Féraud, et N. Voisine (2008). Contact personalization using a score understanding method. *International Joint Conference on Neural Network*, 649–654.

Robnik-Sikonja, M. et I. Kononenko (2008). Explaining classifications for individual instances. *IEEE TKDE 20*(5), 589–600.

Summary

This article presents a new method to explore how the probabilities produced by a classification model may be influenced by a change of the input values. This exploration is here applied to probabilist classifiers. The goal is to increase the predictive probability of a given class by exploring the possible values of the input variables taken independently. The proposed method is presented in a general framework, then detailed for naïve Bayesian classifiers. Its application to data proposed for PAKDD 2007 challenge shows that such approach enables to create useful indicators for sales talks.