Détection par SVM - Application à la Détection de Churn en Téléphonie Mobile Prépayée

Cédric Archaux *,**, Arnaud Martin **, Ali Khenchaf **

* Bouygues Telecom, 20 quai du point du jour, 92100 Boulogne Billancourt
carchaux@bouyguestelecom.fr
http://www.bouyguestelecom.fr

** ENSIETA, 2 rue François Verny, 29806 Brest cedex 9
[archauce, Arnaud.Martin, Ali.Khenchaf]@ensieta.fr
http://www.ensieta.fr/e3i2

Résumé. Dans le cadre de la téléphonie mobile prépayée, les clients ne sont pas engagés contractuellement et peuvent donc cesser leur activité sans préavis. Afin d'estimer l'effort de fidélisation qui peut être engagé au cas par cas, l'opérateur doit donc distinguer les clients qui sont fortement risqués de ceux qui le sont moins. Les machines à support de vecteurs (SVM) sont particulièrement efficaces en classification, et sont particulièrement adaptés aux données bruitées grâce à leur robustesse. Ainsi, nous présentons dans cet article la classification par SVM appliquée à la détection de churn en téléphonie mobile prépayée. Nous montrons que cette approche permet d'obtenir de meilleurs résultats qu'un réseau de neurones.

1 Introduction

Après plusieurs années de très forte croissance, le marché français de la téléphonie mobile a atteint sa maturité et se stabilise. Il devient désormais primordial pour les opérateurs de fidéliser leurs clients afin de prévenir leur fuite à la concurrence (churn). Afin d'évaluer la valeur à terme des clients, la modélisation de la fonction de valeur au cours du temps ainsi que la fonction de survie des clients a été introduite. Les modèles de survie utilisés sont historiquement des modèles paramétriques et des modèles non-paramétriques à risques proportionnels (Rosset et al. 2002). Notre thèse est de présenter ce problème comme par une approche de détection (classification en deux classes) : ceux qui vont cesser leur activité dans un horizon de temps donné et ceux qui vont maintenir leur niveau de consommation.

Les approches déjà étudiées dans le domaine de la téléphonie se sont tournées vers des techniques telles que les chaînes de Markov (Hollmén 2000), les mixtures de gaussiennes et réseaux bayésiens (Taniguchi et al. 1998), les règles d'associations (Rosset et al. 1999), ou encore les réseaux de neurones (Mozer et al. 2002). Dans (Mani et al. 1999), il est montré que l'introduction d'un réseau de neurones multi-couches pour la modélisation de la survie permet d'obtenir de bons résultats en s'appuyant sur une analyse critique d'autres méthodes. Nous avons donc cherché à appliquer cette dernière approche pour la détection du churn. Cependant, nous avons souhaité améliorer la capacité de généralisation du modèle pour tenir compte du volume très important de données que nous devons traiter.

L'approche SVM (Vapnik 1998) tente de séparer des clients à fort risque de fuite des clients moins risqués dans l'ensemble des clients par l'hyperplan optimal qui garantit que

l'écart entre les deux classes soit maximal. Les nouveaux clients pour lesquels nous devons détecter le churn, pourront ainsi ne pas être trop similaires à ceux employés pour trouver l'hyperplan mais être tout de même situés franchement d'un côté ou l'autre de la frontière. La force des SVM tient à leur simplicité de mise en œuvre face à des problèmes difficiles et à des fondements mathématiques solides. Nous avons donc retenu et testé les modèles SVM qui ont fait leur preuves dans d'autres domaines tels que la classification d'images (Goffinet 2001, Laayana 2003), ou la reconnaissance de locuteurs (Gutschoven et al. 2000).

Nous présentons dans un premier temps la méthode de classification par SVM, nous décrivons ensuite plus précisément les bases de données utilisées, puis les résultats de détection des clients à risque obtenus.

2 La méthode SVM

Nous présentons donc dans cette partie comment la méthode SVM s'inscrit dans le cadre de la théorie de l'apprentissage supervisé et comment nous pouvons formaliser le problème de détection de churn à l'aide de cette approche. Nous rappelons enfin le principe des SVM.

2.1 Théorie d'apprentissage supervisé

Soit O un ensemble de clients décrit par un nombre fixe d de caractéristiques (variables descriptives numériques). Prenons un sous-ensemble S de O, l'ensemble de test constitué d'un ensemble de l couples $(x_i, y_i)_{1\le i\le l}$ où x_i est un point de $\mathbf{R}^d$ qui représente les caractéristiques des clients et $y_i=\pm 1$ représente la classe du client x_i (le client est à risque (+1) ou non (-1)). Etant donné les caractéristiques des clients de O-S, l'ensemble d'apprentissage, nous cherchons à estimer si un client de S est à risque ou non, ou encore une estimation de la fonction qui à tout x_i associe un y_i pour l'appliquer à ce nouveau client. Nous cherchons donc la fonction qui réalise la meilleure approximation de la réponse désirée parmi une famille de fonctions $\{f_\alpha\}$ à valeurs dans $\{-1,+1\}$. Les $(x_i, y_i)_{1\le i\le l}$, supposées indépendantes et identiquement distribuées, sont issues d'une distribution de probabilité inconnue $P(x,y)$. Le critère choisi est la minimisation du risque R défini par : $R(\alpha)=\int 1/2\left|y-f_\alpha(x)\right| dP(x,y)$. La probabilité P étant inconnue, R l'est aussi ; par contre, nous pouvons estimer risque empirique sur l'ensemble des observations de la base d'apprentissage : $R_{emp}(\alpha)=1/(2l)\sum_{i=1}^{l}\left|y_i-f_\alpha(x_i)\right|$.

Pour une probabilité au moins égale à 1-η, on a l'inégalité suivante :

$$R(\alpha)\le R_{emp}(\alpha)+\sqrt{1/l(h(\ln(2l/h)+1)-\ln(\eta/4))}\,, \qquad (1)$$

où h est la *VC-dimension* du nom de Vapnik et de Chervonenkis (Guermeur et Paugam-Moisy 1999), c'est le maximum de points pour lesquels les fonctions $\{f_\alpha\}$ associent la bonne classe. Le second terme de la borne supérieure, nommé intervalle de confiance, est une fonction croissante monotone en h. Ainsi, pour h petit, il suffit de minimiser le risque empirique pour minimiser le risque R.

Ainsi, pour garantir une faible valeur de R, nous devons chercher une valeur optimale de la VC-dimension h. C'est un problème de minimisation du risque. Le contrôle du risque consiste donc à contrôler la VC-dimension puisque la taille de l'observation l est

généralement fixée. Vapnik (Vapnik 1998) propose d'appliquer le principe de minimisation du risque structurel dont le but est la minimisation conjointe du risque empirique et de l'intervalle de confiance. En considérant les hyperplans sur $\mathbf{R}^d$ définis par $\{x \in \mathbf{R}^d : x.w+b=0\}$, (Burges 1998) montre que minimiser la VC-dimension revient à minimiser $||w||$.

2.2 Principe des SVM

S'il existe un hyperplan qui sépare les deux classes, les points de l'hyperplan sont décrits par l'équation $x_i.w+b=0$ où w est la normale au plan et $|b|/\|w\|$ la distance entre l'hyperplan et l'origine, voir FIG. 1.

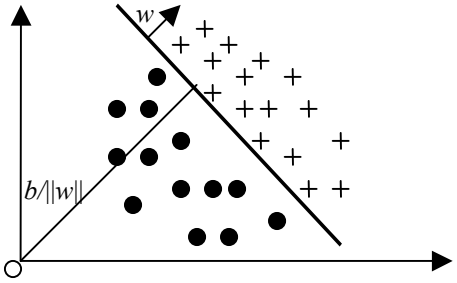

FIG. 1 – *Cas linéairement séparable.*

Soient d_+ (resp. d_-) la distance minimale entre l'hyperplan et la classe des x_i tel que y_i=+1 (resp. y_i =-1). L'hyperplan optimal est celui qui maximise $d_+ + d_- = (1-b)/\|w\| - (-1-b)/\|w\| = 2/\|w\|$. Ceci se traduit par l'existence d'un couple $(w,b) \in \mathbf{R}^d \times \mathbf{R}$ tel que : $x_i.w+b=0$, pour les points de cet hyperplan, avec

$$y_i(x_i.w+b)-1 \geq 0 \text{ , pour tout } i=1,\ldots,l. \quad (2)$$

L'hyperplan optimal est donc déterminé en minimisant $J(w)=\|w\|^2/2$ sous les contraintes (2). Les vecteurs de support sont les points tels que $y_x(x_l.w + b)-1=0$. Il s'agit donc de chercher des constantes w et b vérifiant (2) qui minimisent $J(w)$. Ce système ce résout simplement (Laayana 2003), et montre que pour estimer la classe d'un nouveau client x, on calcule :

$$f(x) = sign((x.w^0)+b^0) = sign(\sum_{i=1}^{l} \alpha_i^0 y_i (x_i.x) + b^0) = sign(\sum_{VS} \alpha_i^0 y_i (x_i.x) + b^0) \quad (3)$$

où VS est l'ensemble des vecteurs de support.

Pour généraliser cette méthode dans le cas où la fonction de décision n'est pas linéaire, l'idée est de plonger les vecteurs d'entrée dans un autre espace de dimension suffisamment grande en utilisant une fonction $\Phi : \mathbf{R}^d \rightarrow H$, tel qu'il existe une fonction K, le noyau :

$$K : \mathbf{R}^d \text{ x } \mathbf{R}^d \rightarrow \mathbf{R}$$
$$(x,x') \rightarrow \Phi(x).\Phi(x').$$

Il suffit donc de chercher l'hyperplan optimal dans l'espace H par la méthode précédente : le couple $(x_i, y_i)_{1 \leq i \leq l}$ est remplacé par $(\Phi(x_i), y_i)_{1 \leq i \leq l}$, en reprenant les formules précédentes, et en utilisant le produit scalaire de H au lieu du produit scalaire de $\mathbf{R}^d$. Enfin, pour estimer la classe d'un client x, il suffit de calculer la fonction : $f(x) = sign(\sum_{VS} \alpha_i^0 y_i K(x_i,x) + b^0)$. Il n'existe cependant pas de méthode ni pour choisir Φ, ni

pour choisir le noyau *K*. Les principaux noyaux utilisés dans les applications sont : les polynômes de degré *p* ($K(x,y)=((x.y)+1)^p$), et les gaussiennes ($K(x,y) = e^{-\|x-y\|^2/2\sigma^2}$).

Nous allons donc appliquer cette approche pour la détection de churn de clients, en comparant les différents noyaux.

3 La base de données

Les bases de données que nous utilisons sont composées de différents types de données :

- des données de facturation comme les montants rechargés par les clients ou les montants prélevés pour les services et options souscrits, ces montants sont des nombres réels qui prennent généralement leurs valeurs dans un ensemble restreint.
- des données relatives aux usages comme le nombre total des appels, la répartition des appels locaux nationaux ou internationaux (pourcentage), la consommation en pic et la consommation moyenne (réels),
- des données relatives à la ligne téléphonique telles que l'ancienneté (entier borné à l'ancienneté de commercialisation des offres prépayées), le plan tarifaire actuel, le nombre de plans tarifaires différents par lequel le client est passé,
- des données relatives aux souscriptions et résiliations de services,
- d'autres informations telles que l'âge ou la catégorie socioprofessionnelle du client, la rentabilité actuelle et la rentabilité précédente, la sélection d'autres options, etc.

A une date de modélisation donnée, nous construisons une base de travail constituée des données relatives à un ensemble *O* de clients actifs (c'est-à-dire qu'ils n'ont pas cessé leur activité). Les clients sont décrits par un nombre fixe *d*=61 de caractéristiques les concernant durant 6 mois écoulés, il s'agit de données caractéristiques des lignes et de données de consommations et rechargements agrégées de façon mensuelle sur chacun des 6 mois de la période d'étude. A ces données est ajouté un indicateur à deux valeurs (+1 et -1) qui indique si les clients ont cessé leur activité durant les trois mois suivant la date de modélisation (+1), ou non (-1). L'ensemble *O* est pris aléatoirement dans la base constituée de l'ensemble des 141000 clients de façon à conserver la proportion de clients qui ont cessé leur activité dans les trois mois.

Pour l'apprentissage et le test du classifieur (en l'occurrence le programme *SvmFu 2.3* développé par le *MIT* et le Cambridge Research Lab de Compaq (Rifkin et al. 2002)) nous avons utilisé la base de travail pour construire trois bases :

- La base d'apprentissage sert à faire apprendre le classifieur, elle caractérise 600 clients (500 clients non churners et 100 clients churners) par *d*=61 caractéristiques.
- Deux bases de tests sont utilisées pour appliquer le classifieur :
  La base de test n°1 est composée également de *l*=600 clients (500 clients non churners et 100 clients churners) ce qui est une volumétrie comparable au fichier d'apprentissage pour tester la qualité de l'apprentissage.
  La base de test n°2 est composée de *l*=2500 clients (2000 clients non churners et 500 clients churners, pour garder la même proportion) volume significativement plus élevé que la base d'apprentissage pour tester la capacité de généralisation du modèle.

4 Résultats

Nous avons appliqué les SVM et le réseau de neurones multicouches sur les mêmes bases d'apprentissage et de tests afin de comparer les résultats. Le tableau TAB 1 présente les taux de bonnes classifications des classifieurs appliqués sur les données de tests :

Taux de bonne prédiction par modèle	SVM noyau polynomial	SVM noyau linéaire	SVM noyau gaussien	Réseau de neurones
Test n°1	44,55 %	87,10 %	88,55 %	87,46 %
Test n°2	45,16 %	80,16 %	84,28 %	81,08 %

TAB 1 - *résultats des classificateurs*

Nous voyons que sur la base de test n°1 les noyaux gaussiens fournissent des résultats comparables à ceux obtenus par le réseau de neurones et le noyau linéaire. En formulant l'hypothèse que l'estimateur suit une loi gaussienne, nous calculons l'intervalle de confiance à 95% du résultat obtenu par le noyau gaussien sur la base de test n°1 : [86,00 ; 91,10], le taux de bonne prédiction du réseau de neurones et du noyau linéaire ne sont pas significativement plus mauvais. Le test sur la base n°2 montre que les résultats des SVM sont meilleurs quand la base de test est d'un volume considérablement supérieur à la base d'apprentissage, ce qui est une propriété que nous recherchons. En effet, le taux de bonne prédiction de churn du réseau de neurones et du SVM à noyau linéaire sont à l'extérieur de l'intervalle de confiance à 95% du noyau gaussien [82,85 ; 85,71]. Les noyaux polynomiaux caractérisent très mal la frontière entre churners et non-churners, leur utilisation dans ce contexte précis ne présente pas d'intérêt.

5 Conclusion et perspectives

Nous avons montré dans cet article que les machines à support de vecteurs peuvent être appliquées au problème de classification des clients en téléphonie mobile prépayée. Nous avons obtenu des résultats significativement meilleurs que l'approche du réseau de neurones multi-couches de (Mani et al. 1999). Un autre intérêt des SVM est la sélection de vecteurs de support grâce auxquels est déterminé l'hyperplan optimal. Les clients employés lors de la recherche de l'hyperplan ne sont alors plus utiles et seuls ces vecteurs supports sont utilisés pour classer un nouveau client. Cela en fait une méthode très rapide.

Ces premiers résultats sont fortement encourageants, et nos travaux vont aller dans le sens de l'intégration de données plus porteuses en information. Il serait intéressant d'étudier l'influence de la taille de la base d'apprentissage. En effet, les SVM permettent généralement d'obtenir de bons résultats avec une taille restreinte pour la base d'apprentissage, ce qui par sa rapidité d'exécution en fait une méthode capable de s'adapter rapidement à la fluctuation du marché. De plus, le modèle de classificateur présenté n'est pas spécifiquement dédié au données temporelles, et c'est un axe sur lequel vont s'orienter nos travaux à venir.

Références

Burges C. (1998), A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery, Vol. 2(2), pp. 121-167, 1998.

Guermeur Y. et Paugam-Moisy H. (1999), Théorie de l'apprentissage de Vapnik et SVM, Support Vector Machines, Apprentissage automatique, Hermes Sciences Publications, Paris 1999.

Goffinet J. (2001), Machines à vecteur de support pour la détection et le suivi de personnes sur des séquences vidéo, rapport de stage, juillet 2001.

Gutschoven B. et Verlinde P. (2000), Multi-modal Identity Verification using Support Vector Machines (SVM), International Conference on Information Fusion, Paris, France, 10-13 juillet 2000.

Hollmén J. (2000), User Profiling and Classification for Fraud Detection. Thèse de doctorat, University of Helsinki, 2000.

Laayana H. (2003), Détection par SVM – Application à la détection de roches pour le recalage d'images sonar, rapport de DESA, juillet 2003.

Mani D.R., Drew J., Betz A., Datta P. (1999), Statistics and data mining techniques for lifetime value modeling, Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 94-103, 1999.

Mozer M.C., Dodier R., Colagrosso M.D., Guerra-Salcedo C., Wolniewicz R. (2002), Prodding the ROC Curve: Constrained Optimization of Classifier Performance, Advances in Neural Information Processing Systems 14, MIT Press, 2002.

Rifkin R., Moreno P., Nicpanski H., Alvira M., Paris J., King V., Nadermann M., SvmFu Documentation, (2002), http://five-percent-nation.mit.edu/SvmFu/.

Rosset S., Murad U., Neumann E., Idan Y., Pinkas G. (1999), Discovery of fraud rules for telecommunications-challenges and solutions, Proceedings ACM SIGKDD, 1999.

Rosset S., Neumann E., Eick U., Vatnik N., Idan Y. (2002), Customer lifetime value modeling and its use for customer retention planning, Proceedings of the eighth ACM SIGKDD, pp. 332-340, 2002.

Taniguchi M., Haft M., Hollmén J., Tresp V. (1998), Fraud detection in communications networks using neural and probabilistic methods, ICCASP, Vol 2, pp 1241-1244 1998.

Vapnik V. (1998), Statistical Learning Theory. John Wiley & Sons, 1998.

Veropoulos, K., Campbell, C. et Cristianini, N. (1999), Controlling the Sensitivity of Support Vector Machines. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI99). Stockholm, Sweden, 1999.

Summary

Within the context of prepaid mobile telephony, customers are not linked to their operator by contract and thus can cease their activity without notice. In order to estimate the fidelization efforts which can be engaged towards each individual customer, the operator must distinguish the customers presenting a strong churn risk from those less risky. Thanks to their robustness, Support Vectors Machines (SVM) are particularly effective in classification and adapted to noisy data. Thus, the objective of this article is to present the application of SVM classification to churn detection in prepaid cellular telephony. We show that this approach gives better results than neural networks.