An SVM based Churn Detector in Prepaid Mobile Telephony

Cédric Archaux [1, 2], Hicham Laanaya [2], Arnaud Martin [2], Ali Khenchaf [2]

[1] *Bouygues Telecom, 20 quai du point du jour, 92640 Boulogne Billancourt, France*
[2] *Laboratoire E³I², ENSIETA, 2 rue François Verny, 29806 Brest cedex 9, France*
carchaux@bouyguestelecom.fr
[Cedric.Archaux, Hicham.Laanaya, Arnaud.Martin, Ali.Khenchaf]@ensieta.fr

Abstract

The context of prepaid mobile telephony is specific in the way that customers are not contractually linked to their operator and thus can cease their activity without notice. In order to estimate the retention efforts which can be engaged towards each individual customer, the operator must distinguish the customers presenting a strong churn risk from the other. This paper presents a data mining application leading to a churn detector. We compare Artificial Neural Networks (ANN) which have been historically applied to this problem, to Support Vectors Machines (SVM) which are particularly effective in classification and adapted to noisy data. Thus, the objective of this article is to compare the application of SVM and ANN to churn detection in prepaid cellular telephony. We show that SVM gives better results than ANN on this specific problem.

1. Introduction

The study of survival functions was historically introduced by [2] and [6] who worked on parametrical and non parametrical models of patients survival probabilities. [11] shown that this approach could be applied for modeling the mobile telephones customers survival functions.

Our goal in this paper is to solve a detection problem: we want to separate the customers who will cease their activity in a given horizon from those who will maintain their activity. The approaches already studied in the field of telephony were techniques such as Hidden Markov Models [5], the Gaussian mixtures and bayesian networks [12], association rules [10], and neural networks [9]. [8], shown that multi-layer perceptrons obtained good results on the survival function modeling. We have experimented this last model on our problem of churn detection which is somewhat different from the previous application. However, we wanted to improve the generalization capacity of the model in order to apply it to large volume of data.

The SVM approach introduced by [13] tries to separate the risky customers from the other customers by an optimal hyperplane which guarantees that the margin between the two classes is maximum. The new customers we are to detect, could thus not be too similar with those used to find the hyperplane but to be located frankly on one side or the other of the border. The force of the SVM is to simply face difficult problems thanks to solid mathematical bases.

We have thus retained and tested both SVM and ANN on the same database. We compare in this article polynomial, linear and gaussian kernels to multi-layer perceptron using different transfer functions. We compare the models on the rate of good classification and robustness on two different sized test databases.

Our work follows the classical Knowledge Discovery in Databases Process introduced in [3] and described in Fig. 1.

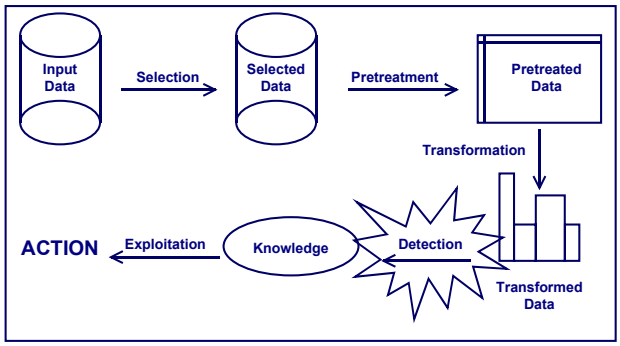

Fig. 1: The Knowledge Discovery in Database methodology.

This paper begins with the description of the data we are working on and our pretreatment and transformation step. Then we present the two detection models and give results of the good classification rate on the different test bases and conclude.

2. Data Bases

The databases we use in this study are composed of different types of data:

- invoicing data such as the amounts refilled by the clients or amounts withdrawn for the services and the options subscribed, these amounts are real numbers which values are generally in a restricted set,
- data relating to the uses such as the total numbers of calls, the share of the local, national or international

calls (percentage), the consumption's peaks and average consumption (real),

- data relating to the telephony line such as the age (whole number limited to the date of the commercialization of prepaid offers), the current tariff plan, the number of different plans the client passed by,
- data relating to subscriptions and cancellation of services,
- other information such as age or socio-professional category of the client, the current profitability and the previous profitability, the selection of others options, etc.

Given a modelization date, we build a learning base constituted by data relating to a set O of active clients (*i.e.* clients who have not ceased their activity). The clients are described by a fixed number of characteristics d=61 relative to them for the last 6 months. Those data are characteristics of the lines and consumption and refilling data monthly aggregated on each of the 6 months of the learning period.

We add to these data a two valued indicator (+1 and -1) which show whether the clients ceased their activity during the 3 months following the modelization date (+1), or not (-1). The learning sets O are taken from a 141.000 lines database constituted, in order to keep the proportion of clients that cease their activity during the 3 months. We take several sample sizes and churner ratios as learning and testing samples those samples are detailed in the results section.

3. SVM method

In this part, we present how the SVM method fits with the supervised learning theory and how we can formalize the problem of churn detection with the help of this approach. Finally, we remind of the SVM's principles.

3.1. Supervised learning theory

Let O be a set of clients described by a fixed number of characteristics d (numerical descriptive variable). Let S be a subset of O, the test set is made up of a set of l couples $(x_i, y_i)_{1\le i\le l}$ where x_i is a point of $\mathbf{R}^d$ which represents the clients' characteristics and $y_i=\pm 1$ represents the x_i client's class (risky client (+1) or non risky client (-1)). Considering the O-S clients' characteristics, the learning set, we try to estimate if a client of S is risky or not, or an estimation of the function in which every x_i is linked with an y_i in order to apply it to a new client. So, we are looking for the function that realize the best approximation of the wanted answer among a group of functions $\{f_\alpha\}$ with values in $\{-1,+1\}$. The $(x_i, y_i)_{1\le i\le l}$, supposed to be independent and identically distributed, are from an unknown probability $P(x,y)$. The selected criterion is the minimization of the risk R defined by: $R(\alpha)=\int 1/2\left|y-f_\alpha(x)\right| dP(x,y)$. Considering that the probability P is unknown, so is R ; on the contrary, we an estimate empiric risk on the whole observations of the learning base:

$$R_{emp}(\alpha)=1/(2l)\sum_{i=1}^{l}\left|y_i-f_\alpha(x_i)\right|.$$

For a probability at least equal to 1-η, $0\le\eta\le 1$, we have the following inequality:

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

where h is the *VC-dimension* in name of Vapnik and Chervonenkis [13], this is the maximum of points for which the functions $\{f_\alpha\}$ combine the right class. The second term of the upper limit, called confidence interval, is a monotonous increasing function on h. Thus, for small h, minimize the empiric risk is enough to minimize the risk R.

So, in order to guarantee a weak value of R, we have to look for an optimum value of the VC-dimension h. It is as problem of risk minimization. The control of the risk consists in controlling the VC-dimension because the size of the observation l is generally fixed. Vapnik in [13] suggests to apply the principle of minimization of the structural risk which objective is the joint minimization of the empiric risk and of the confidence interval. Considering the hyperplanes on $\mathbf{R}^d$ defined by $\{x\in\mathbf{R}^d : x.w+b=0\}$, Burges in [1] shows that minimizing the VC-dimension comes down to minimize $\|w\|$.

3.2. SVM principle

If an hyperplane which split the two classes exist, the points of the hyperplane are described by the equation $x_i.w+b=0$ where w is the normal to the plan and $|b|/\|w\|$ is the distance between the hyperplane and the origin. See FIG. 2.

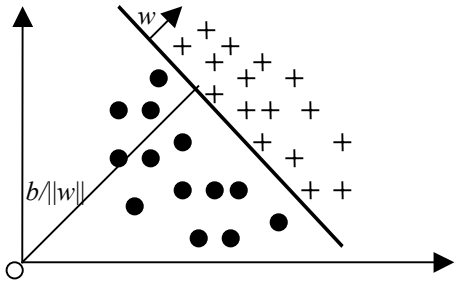

FIG. 2 – *Linearly separable case.*

Let d_+ (resp. d_-) be the minimum distance between the hyperplane and the class of x_i so that $y_i=+1$ (resp. $y_i=-1$). The optimal hyperplane is the one that maximize $d_+ + d_- = (1-b)/\|w\| - (-1-b)/\|w\| = 2/\|w\|$. This is rendering by the existence of a couple $(w,b)\in\mathbf{R}^d\times\mathbf{R}$ such as: $x_i.w+b=0$, for the points of the hyperplane, with

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

So, the optimal hyperplane is determined minimizing $J(w)=\|w\|^2/2$ under the constraints (2). Support vectors are points such as $y_i(x_l.w+b)-1=0$. It is a question of

finding the constants w and b that confirm (2) which minimize $J(w)$. This system is simply resolved [7], and shows that in order to estimate the class of an other client x, we calculate:

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

where VS is the set of the support vectors.

To generalize this method in case of the function of decision is not linear, we have to plunge the entry vectors in an other space large enough using a function $\Phi : \mathbf{R}^d \rightarrow H$, so can exist a kernel function K:

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

It is enough looking for the optimal hyperplane in the space H by the previous method : the couple $(x_i, y_i)_{1 \le i \le l}$ is switched by $(\Phi(x_i), y_i)_{1 \le i \le l}$, repeating the previous formulas, and using the scalar product of H instead of the scalar product of $\mathbf{R}^d$. Finally, to estimate the class of a client x, calculate the following function is enough : $f(x) = sign(\sum_{VS} \alpha_i^0 y_i K(x_i, x) + b^0)$. However, it does not exist any method to choice Φ, nor to choice the kernel K. The main kernels used in the applications are: the polynomial of degree p ($K(x,y)=((x.y)+1)^p$), and the gaussian ($K(x,y) = e^{-\|x-y\|^2 / 2\sigma^2}$).

So, we will apply this approach to detect the churn of clients, comparing the different kernels.

3.3. SVM Results

Our first goal is to compare the results of the polynomial kernels, we train 6 different polynomials kernels detectors from first to sixth degree. We compare the detectors by the rate of good classification on the test base constituted of 20 percents of churners. We notice in FIG. 3 that fourth level polynoms give better results.

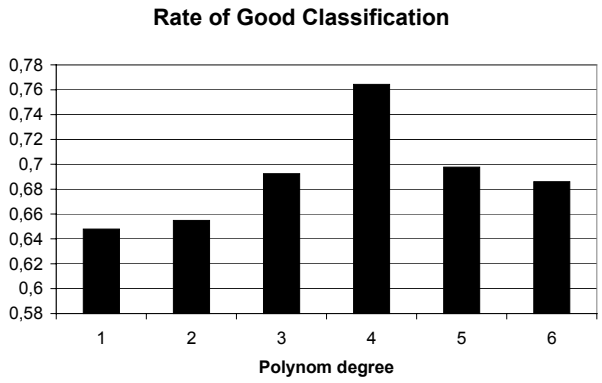

FIG. 3. Performance of polynomial kernels

We then compare several gaussian kernels, we tested the radius values from 0.1 to 0.5 and notice in FIG. 4. that we obtained the best good classification rate with a 0.3 radius.

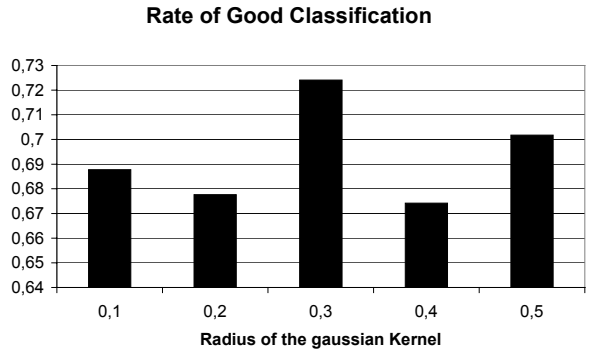

FIG. 4. Performance of gaussian kernels

We then wanted to measure the impact of the learning base size, and we applied the gaussian kernel (radius 0.3) on 10 different learning sets made scaling from 1 000 to 10 000 lines. We notice that after 6 000 lines there is no increase in the good classification rate.

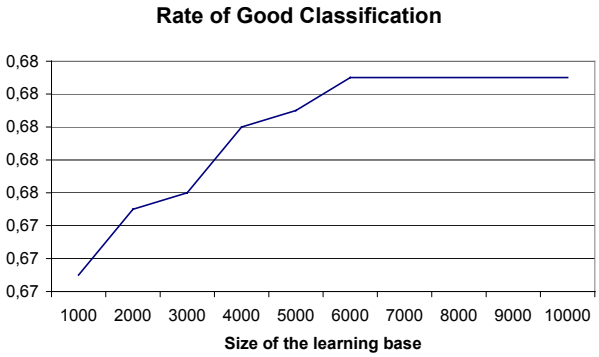

FIG. 5. Impact of the learning size

With the 6 000 lines learning set we wanted to measure the impact of the ratio of churners in the learning set on the good classification rate, so we built 5 different 6 000 lines sets made of respectively 17, 33, 50, 67 and 83% of churners. We learnt models on this five learning sets and measured their good classification rates on 2 test bases made of 6 000 and 60 000 lines.

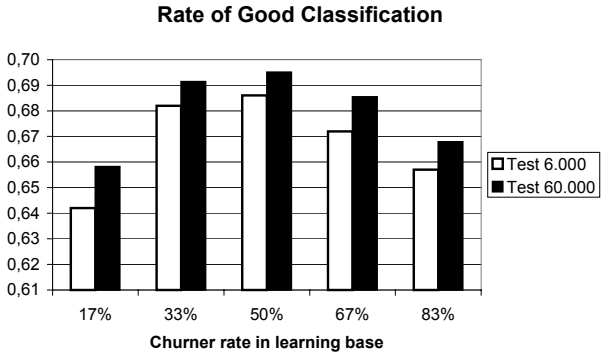

FIG. 6. Impact of the churner rate in the learning base

The immediate conclusion of Fig. 6 is that we obtain the best model with 50% of churners in the learning base. As a conclusion to the SVM results we obtain the best

results with a 6000 lines learning base constituted of 50 percents of churners.

4. Neural Networks

We apply in this part a classical multilayer perceptron. on the same learning and test bases. After having compared it to the Levenberg-Marquadt and standard backpropagation the learning method we retain is the conjugate gradient. The structure of the perceptron is one hidden layer, we increase the number of neurons of this layer until there is a stabilization in the results. The best rate of good classification we obtain is 69,6 percents on the 6000 lines test base. TAB 1 shows the rate of good classifications of the classifiers applied on the data of tests:

	6000	60000
linear	68,6%	69,5%
gaussian	68,5%	69,2%
polynomial	69,8%	69,8%
ANN	69,60%	67,20%

TAB 1 – *results of the classifiers*

We see that on the test base with 6 000 lines the polynomial kernel gives results comparable to the ones obtained by the neural network. The rate of good prediction of the neural network and the svm-based detectors are not significantly different on the 6.000 lines test base. The test on 60 000 lines test base shows that the results of the SVM are better when the base has a volume significantly higher to the learning base, which a property that we are looking for. The rate of prediction of the neural network is out of the confidence interval at 95% of the polynomial kernel. Thus the result of the polynomial kernel based SVM detector is significantly better than those obtained with the neural network.

5. Conclusion and prospects

We have shown in this article that support vectors machines can be applied to customer classification in prepaid mobile telephony. We have obtained with SVM significantly better results than with the approach of the multilayer neural networks. An other interest is the selection of support vectors which determines the optimal hyperplane. The clients used for during the research of the hyperplane are no longer useful and only the support vectors are used to classify a new client. That is why this method is rapid.

These first results are highly promising, and our research will go in the way to integrate data with more information. SVM usually provide good results in a context of classification of data composed of few lines and lots of columns. When applied on our data, SVM techniques gives good results with a small learning base. Thanks to its quick application and robustness we could face noisy data and refresh frequently the results which enables us to quickly adapt to the market evolution. The presented detector is not specifically dedicated to temporal data, the temporal information contained in the data is not fully exploited. Our prospects are to include this classifier to an integrated system dedicated Customer Relationship Management. Our work will also be oriented to the amelioration of the data preparation and the construction of more temporal behavior indicators.

6. References

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

[2] Cox D.R. "Regression models and life tables". Journal of the Royal Statistical Society, B34: pages 187-220 , 1972.

[3] Fayyad U., Piatetsky-Shapiro G. and Smyth. P. "The KDD process for extracting useful knowledge from volumes of data". In Communications of the ACM archive, Volume 39, Issue 11, pages 27-34 , 1996.

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

[5] Hollmén J., "User Profiling and Classification for Fraud Detection". PhD Thesis, University of Helsinki, 2000.

[6] Kaplan E.L. and Meier R. "Nonparametric Estimation From Incomplete Observations" Journal of the American Statistical Association, pages 457-481 , 1958.

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

[8] Mani D.R., Drew J., Betz A., and Datta P., "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.

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

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

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

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

[13] Vapnik V., "Statistical Learning Theory". John Wiley & Sons, 1998.