

Predicting Churners in Telecommunication Using Variants of Support Vector Machine

Sindhu M E¹ and Vijaya M S²

¹*Mphil Research Scholar, Department of Computer Science, PSGR Krishnammal College for Women, Coimbatore-641004, Tamilnadu, India.*

²*Associate Professor, PSGR Krishnammal College for Women, Coimbatore-641004, Tamilnadu, India.*

ABSTRACT : Customer acquisition and customer retention are one of the most competitive factors in most of the companies. Due to ever increasing competitions of customers in companies, the company owners are unable to maintain the customer satisfaction which leads to customer churn. The customer wishes to leave the service of the company causes churn. Most of the sectors are affected by churn problems. Telecommunication is one of the main industries that are affected by churn problem. Prediction of customers who are at risk of leaving a company is known as churn prediction and it is imperative for sustainable growth of a company. Supervised classification technique suits best for solving this problem. This research work employs the variants of Support Vector Machine such as Proximal Support Vector Machine, Active Support Vector Machine and Lagrangian Support Vector Machine for creating the prediction models and the best model is recognized based on predictive accuracy.

KEYWORDS : Churn prediction, customer churn, churner, non-churner and customer acquisition.

I. INTRODUCTION

In the world of ever increasing competition on the telecommunication, companies have become observant that they should put much effort not only to convince customers but also to keep hold of existing customers. Churners are persons who quit a company's service for certain reasons. One of the major reasons for predicting churn is that it costs less to retain existing customers than to acquire new customers. Churn prediction methods gives the prediction about customers who likely to churn in the near future whereas churn management helps on the other side which aims to identify such churners and to carry out some positive actions to minimize the churn effect. Churn prediction is a binary classification task which differentiates churners and non-churners. The concept of customer relationship management has gained its importance in marketing domain. Many previous CRM-related researches have used data mining techniques to analyze and understand customer behaviors and characteristics. In general, churn means to role of the customers who are about to move their usage of service to a competing other provider. There are two basic categories of churners, voluntary and involuntary. Voluntary churn occurs when the customer initiates termination of the service contract. Involuntary churners are the customers that the telecommunication company decides to remove from the subscribers list.

II. LITERATURE SURVEY

Many prediction algorithms have been proposed so far to predict the customer churn. In the research work [1], the author applied the well known data mining methodology CRISP-DM (cross-industry standard process for data mining) to investigate network usage behaviors of the ISP (Internet service providers) subscribers in Taiwan. They used Attribute-Oriented Induction (AOI) method for discovering characteristics and discrimination knowledge of ISP customers from the ISP traffic data. Authors in [2] proposed three different techniques which are decision tree, support vector machine, and neural network for classification. K-mean techniques for clustering. A system developed in [3] focused genetic programming approach to predict churn and proposed the intelligent churn prediction technology to develop a new hybrid model to improve the performance and accuracy for churn prediction.

Also SVM and K-mean cluster was performed. In [4], the authors proposed various classification techniques to find number of features subset in different size and dimension. The experiments were carried out using decision tree J48, C5.0 classification. In this research work [5], the churn prediction model has been proposed by various classification techniques like decision tree, logistic regression and neural network. They have mainly focused on neural network. The neural network was modeled to get an insight about the importance of three types of features in the demographic, billing and usage features. Authors in [6] proposed new subset of features in order to improve the accuracy of customer churn prediction in the wireless telephony industry. The new features were categorized as contract-related features, call patterns description features, and calls pattern changes description features. To evaluate the features, experiments based on two probabilistic data mining algorithms Naïve Bayes and Bayesian Network, were performed their results are compared to those obtained from using C4.5 decision tree, a widely used algorithm in many classification and prediction tasks. In existing research work, various researchers proposed classification technique like decision tree, support vector machine, neural network. From the literature it was well appreciated that Support Vector Machine is a powerful classification technique with high generalization power than other classification algorithms. Hence in this research work the variants of support vector machine proposed by Olvi Mangasarian [11] have been used to automate the prediction of churners. The proposed methodology models the churn prediction as a binary classification task and the classification task is implemented using variants of SVMs such as PSVM, ASVM and LSVM. The prediction model has been trained using training dataset and the trained model is built. Finally the trained model is used to predict whether the customer is churner or non-churner. Various stages of the proposed implementation are described in section IV.

III. SUPPORT VECTOR MACHINE

Support vector machines (SVM) are learning systems that use a hypothesis space of linear functions in a high dimensional feature space, trained with a learning algorithm from optimization theory that implements a learning bias derived from statistical learning theory. The support vector machine is a training algorithm for learning classification and regression rules from data. The geometrical interpretation of support vector classification is that the algorithm searches for the optimal separating surface, i.e. the hyper plane that is equidistant from the two classes. This can be extended to multi-class problems. Kernel functions are then introduced in order to construct non-linear decision surfaces.

PROXIMAL SUPPORT VECTOR MACHINE: A simpler classifier called proximal support vector machine was recently implemented wherein each class of points is assigned to the closest of two parallel planes in input and feature space that are pushed apart as far as possible. This formulation, leads to an extremely fast and simple algorithm for generating a linear or non linear classifier that is obtained by solving a single system of linear equations.

In PSVM the point of departure is that, the optimization problem of standard linear SVM is replaced by the following problem subject to the constraints

$$\min_{(\mathbf{w}, \gamma, \mathbf{y})} \frac{1}{2} \|\mathbf{y}\|^2 + \frac{1}{2} (\mathbf{w}^T \mathbf{w} + \gamma^2)$$

$$D(A\mathbf{w} - e\gamma) + \mathbf{y} = \mathbf{e}$$

There is no non-negativity constraint on \mathbf{y} , because now ' \mathbf{y} ' represents deviation of the point from the plane passing through the centroid of the data cluster (A_+ or A_-) of which the point belongs. The Karush-Kuhn-Tucker (KKT) necessary and sufficient optimality conditions for this equality constrained optimization problem are obtained by setting equal to zero the gradients with respect to \mathbf{w} , γ , \mathbf{y} , \mathbf{u} of the Lagrangian:

$$L(\mathbf{w}, \gamma, \mathbf{y}, \mathbf{u}) = \frac{1}{2} \mathbf{y}^T \mathbf{y} + \frac{1}{2} (\mathbf{w}^T \mathbf{w} + \gamma^2) - \mathbf{u}^T [D(A\mathbf{w} - e\gamma) + \mathbf{y} - \mathbf{e}]$$

Where \mathbf{u} is the lagrangian multipliers associated with equality constraints. From the first three equations we have

$$\mathbf{w} = A^T D\mathbf{u} \qquad \gamma = -e^T D\mathbf{u}$$

Substituting w, γ, y in the fourth equation we get u in terms of the given data D and A .

$$\begin{aligned}\frac{\partial L}{\partial w} &= w - A^T D u = 0 \\ \frac{\partial L}{\partial r} &= \gamma + e^T D u = 0 \\ \frac{\partial L}{\partial y} &= v y - u = 0 \\ \frac{\partial L}{\partial u} &= D(Aw - e\gamma) + y - e = 0 \\ [D(AA^T - ee^T)D]u + \frac{u}{v} &= e \\ D(AA^T D u - ee^T D u) + \frac{u}{v} &= e \\ [D(AA^T - ee^T)D + \frac{I}{v}]u &= e \\ u &= [\frac{I}{v} + D(AA^T - ee^T)D]^{-1} e \\ u &= [\frac{I}{v} + HH^T]^{-1} e\end{aligned}$$

where $H = D[A - e]$. The explicit solution for w, γ, y can be obtained using the above u . But it requires inversion of possibly massive $m \times m$ matrix. To overcome this, Sherman-Morrison-Woodbury formula for matrix inversion is used, which results in which requires only inversion of

$$\mathbf{u} = v(\mathbf{I} - \mathbf{H}(\frac{\mathbf{I}}{v} + \mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T) \mathbf{e}$$

the dimension of which is only $(n+1) \times (n+1)$, n is the number of predictor variables. Once u is obtained w and γ , can be obtained from the corresponding equations. The classification of new instance is based on the following

$$w^T x - \gamma > 0 \text{ then } x \in A^+$$

$$w^T x - \gamma < 0 \text{ then } x \in A^-$$

$$\gamma + e^T D t = 0$$

$$v y - t = 0$$

$$D(KD u - e\gamma) + y = e$$

LAGRANGIAN SUPPORT VECTOR MACHINE: A fast and extremely simple algorithm, LSVM, has the ability to handle massive problems. Lagrangian support vector machine uses Karush Khun Tucker optimality condition for solving optimization problem.

The problem is to find u using Karush-Kuhn-Tucker necessary and sufficient optimality conditions are

$$\min f(u) = \frac{1}{2} u^T Q u - e^T u$$

$$u \geq 0$$

$$0 \leq u \perp Qu - e \geq 0$$

For any a and b two real vectors, we have

Thus the iterative scheme which constitutes LSVM algorithm is

$$\begin{aligned} 0 \leq a \perp b \geq 0 &\Leftrightarrow a = (a - \alpha b)_+ \alpha > 0 \\ u^{i+1} &= Q^{-1}[e + ((Qu^i - e) - \alpha u^i)_+] \\ i &= 0, 1, 2, \dots \end{aligned}$$

Algorithm converges linearly if: $0 < \alpha < 2/v$. In practice α is taken as $1.9/v$. This Fast Newton iterative method which requires the inversion of single matrix of the order of the input space plus one.

ACTIVE SUPPORT VECTOR MACHINE: Active support vector machine algorithm proposes an iterative method for determining the dual variable u . This method partitions the variable u into basic and non-basic variables. The non basic variables are those which are set to zero. The values of the basic variables are determined by finding the gradient of the objective function with respect to these variables, setting equal to zero and solving the resulting linear equations for these basic variables. If any basic variable takes on a negative value after solving linear equations, then it is set to zero and becomes non basic. Consider the following standard linear SVM with control parameter $v > 0$

$$\begin{aligned} \frac{\partial L}{\partial w} &= w - A^T D u = 0 \\ \frac{\partial L}{\partial r} &= \gamma + e^T D u = 0 \\ \frac{\partial L}{\partial y} &= v y - u = 0 \\ \frac{\partial L}{\partial u} &= D(Aw - e\gamma) + y - e = 0 \\ L(w, \gamma, y, u) &= \frac{v}{2} y^T y + \frac{1}{2} (w^T w + \gamma^2) - u^T [D(Ax - e\gamma) + y - e] \end{aligned}$$

Lagrangian of this problem is given by

$$\begin{aligned} \text{Minimize} \quad & v e^T y + \frac{1}{2} w^T w \\ \text{subject to} \quad & D(Ax - e\gamma) + y \geq e \end{aligned}$$

Substituting w, γ, y in the fourth equation we get u as below

$$\begin{aligned} D(AA^T D u - e e^T D u) + \frac{u}{v} &= e \\ [D(AA^T - e e^T) D] u + \frac{u}{v} &= e \\ [D(AA^T - e e^T) D + \frac{I}{v}] u &= e \\ u &= \left[\frac{I}{v} + D(AA^T - e e^T) D \right]^{-1} e \\ u &= \left[\frac{I}{v} + H H^T \right]^{-1} e \end{aligned}$$

where $H = D[A \quad -e]$. By applying Sherman Morrison Woodbury formula

$$\mathbf{u} = \mathbf{v}(\mathbf{I} - \mathbf{H}(\frac{\mathbf{I}}{\mathbf{v}} + \mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T)\mathbf{e}$$

which requires the inversion of matrix of dimension only $(n+1) \times (n+1)$, n is the number of predictor variables instead of $m \times m$ matrix. Substituting back in the Lagrangian function and simplifying we get the following dual function in terms of u .

$$\min f(u) = \frac{1}{2}u^T Qu - e^T u$$

$$u \geq 0$$

IV. EXPERIMENT AND RESULTS

Three experiments have been carried out by implementing the variants of SVMs such as PSVM, ASVM and LSVM. These SVMs are implemented in matlab. The data are collected from a private telecommunication company. Since the churn prediction task is modeled as binary classification the class label +1 and -1 are assigned for churners and non-churners respectively. The SVM based models are built and the performance of trained models is evaluated using 10-fold cross validation for its predictive accuracy, learning time and number of support vectors.

DATA ACQUISITION: Data acquisition is an important task in any prediction problem. The information about 750 customers is collected from a private telecommunication company for the churn prediction implementation. 23 different attributes pertaining to customer entities such as customer id, customer name, area, state, account length, area code, phone number, international plan, voice mail plan, number of voice mail messages, total day minutes, total day calls, total day charge, total evening minutes, total evening calls, total evening charge, total night minutes, total night calls, total night charge, total international minutes, total international calls, total international charge, number of customer service calls etc., are gathered. The high ranked attributes are selected using information gain attribute evaluation feature selection method for improving the performance of the classification.

CLASSIFICATION USING PSVM: The regularization parameter C is assigned values ranging from 0.1 to 1. It is observed that the training was stabilized for $C = 0.4$. The results of PSVM based prediction model are shown in Table I.

Table I. Results of PSVM

Parameter C	Prediction Accuracy	Learning Time (in secs)	Number of Support Vectors
0.1	99.34	0.56	325
0.2	98.67	0.67	423
0.3	99.42	0.51	321
0.4	100	0.34	463
0.5	99.78	0.65	221
0.6	98.34	0.43	452
0.7	97.23	0.55	242
0.8	99.42	0.46	419
0.9	99.1	0.54	190
1	99.23	0.64	460

CLASSIFICATION USING LSVM: The regularization parameter C is assigned values ranging from 0.1 to 1. It is observed that the training was stabilized for $C = 0.5$. The results of LSVM based prediction model are shown in Table II.

Table II. Results of LSVM

Parameter C	Prediction Accuracy	Learning Time (in secs)	Number of Support Vectors
0.1	84.01	0.76	219
0.2	84.19	0.87	322
0.3	84.23	0.51	210
0.4	84.4	0.54	176
0.5	85.5	0.43	228
0.6	85.52	0.43	201
0.7	85	0.32	212
0.8	85.46	0.48	139
0.9	85.39	0.72	107
1	85.42	0.61	216

CLASSIFICATION USING ASVM: The regularization parameter C is assigned values ranging from 0.1 to 1. It is observed that the training was stabilized for C = 0.4. The results of ASVM based churn prediction model are shown in Table III.

Table III. Results of ASVM

Parameter C	Prediction Accuracy	Learning Time (in secs)	Number of Support Vectors
0.1	86.41	0.90	258
0.2	86	0.56	321
0.3	86.16	0.69	299
0.4	86.25	0.54	362
0.5	85.01	0.78	357
0.6	86.21	0.88	219
0.7	84.15	0.87	356
0.8	84.30	0.89	289
0.9	85.45	0.76	331
1	83.46	0.79	304

COMPARATIVE ANALYSIS: The comparative analysis of all three experiments has been carried out and the comparative results indicate that PSVM based classification model yields a better performance when compared to other models. The comparative results of PSVM, LSVM and ASVM in terms of accuracy, learning time and average number of support vectors are shown in Table IV and illustrated in Figure.1, Figure.2 and Figure.3.

Table IV – Comparison of accuracy, learning time and number of support vectors

Classifier	PSVM	LSVM	ASVM
Accuracy (%)	100	85.5	86.25
Learning time (in secs)	0.34	0.43	0.54
Average number of support vectors	352	203	309

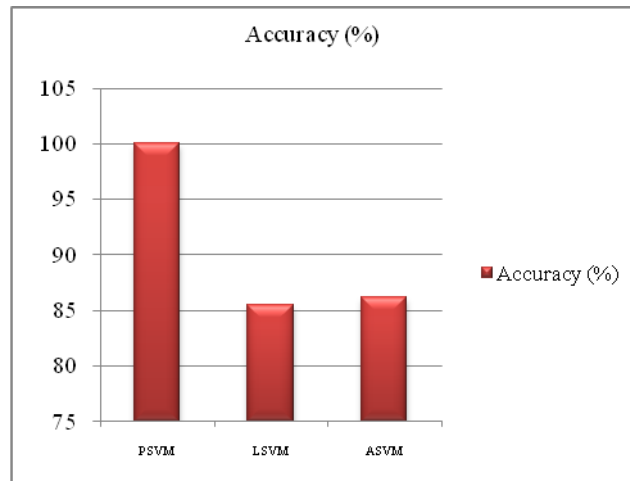


Figure 1: Accuracy of various SVMs

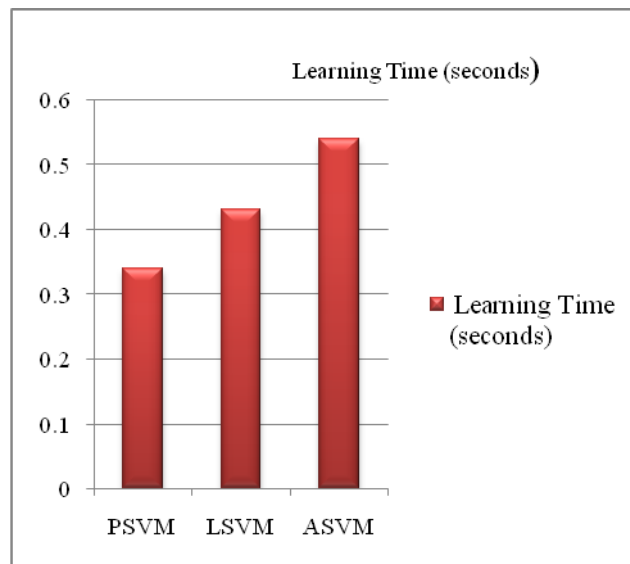


Figure 2: Learning time of various SVMs

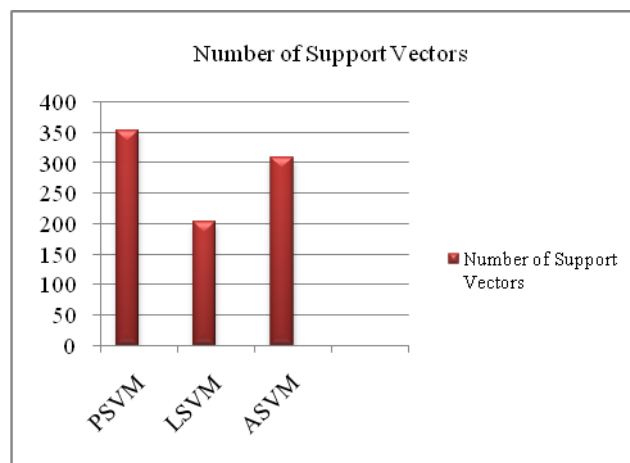


Figure 3: Number of support vectors of various SVMs

From the comparative analysis, it is observed that the number of support vectors in PSVM model is high, which consumes more memory. But the PSVM model is more efficient in terms of predictive accuracy and

computational time. Since predictive accuracy plays the vital role in real time predictions, the PSVM model can be appropriately used in churn prediction and can be implemented in high end systems thus compromising the memory.

V. CONCLUSION

This paper, demonstrates the application of various SVMs in modeling churn prediction as binary classification task. The proximal support vector machine, lagrangian support vector machine and active support vector machine were employed in building the predictive models. Performance of the learned models was evaluated based on their predictive accuracy, learning time and number of support vectors. The most efficient model comparing to the existing churn prediction models was found to be PSVM based churn prediction model and was recommended for predicting whether the telecom customer is churner or non-churner.

REFERENCES

- [1] Shen –Tun Li, Shu Ching Kuo “Knowledge discovery in financial investment for forecasting and trading strategy through wavelet-based SOM networks”, Expert Systems with applications, 34(2): 935-951.
- [2] Essam Shaaban, Yehia Helmy, Ayman Khedr, Mona Nasr, “ A Proposed Churn Prediction Model” International Journal of Engineering Research and Applications (IJERA) SSN: 2248-9622 www.ijera.com Vol. 2, Issue 4, June-July 2012, pp.693-697.
- [3] Imran Khan, Imran Usman and Tariq Usman, “Intelligent Churn prediction for Telecommunication Industry”, International Journal of Innovation and Applied Studies ISSN 2028-9324 Vol. 4 No. 1 Sep. 2013, pp. 165-170, 2013.
- [4] Kamalraj N and Malathi A , “Applying Data Mining Techniques in Telecom Churn Prediction”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 10, October 2013.
- [5] Afaq Alam Khan and Sanjay Jamwal, “Applying Data Mining to Customer Churn Prediction in an Internet Service Provider”, International Journal of Computer Applications (0975 – 8887) Volume 9– No.7, November 2010.
- [6] Clement Kirui, Li Hong, Wilson Cheruiyot and Hillary Kirui, “Predicting Customer Churn in Mobile Telephony Industry Using Probabilistic Classifiers in Data Mining”, IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 2, No 1, March 2013.
- [7] Dejan Radosavljevik, Peter van der Putten, Kim Kylesbech Larsen, “The Impact of Experimental Setup in Prepaid Churn Prediction for Mobile Telecommunications: What to Predict, for Whom and Does the Customer Experience Matter”, Transactions on Machine Learning and Data Mining, Vol. 3, No. 2 pp. 80-99, 2010.
- [8] Den Poel D V and Lariviere B, “Customer attrition analysis for financial services using proportional hazard models”, European Journal of Operational Research, 157(1):196{217, 2004}
- [9] Camilovic D, “Data mining and CRM in telecommunications”, Serbian Journal of Management 3 (1) (2008) 61 – 72.
- [10] Kim H S and Yoon C H “Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market” Telecommunication Policy 28(6): 751-765. (2004).
- [11] Olvi L. Mangasarian and David R. Musicant, Lagrangian Support Vector Machines, Journal of Machine Learning Research, 1:161-177, 2001. <ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/01-02.ps>.
- [12] U. Devi Prasad and S. Madhavi, “Prediction of Churn Behaviour of Bank Customers Using Data Mining Tools”, Business Intelligence Journal Vol.5 No.1, ISSN: 1918-2325, January – 2012.
- [13] K. Coussement and D. V. Poel, “Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques,” Expert Syst. Appl., vol. 34, no. 1, pp.313–327, Jan. 2008.
- [14] S. Neslin, S. Gupta, W. Kamakura, J. Lu, and C. Mason, “Detection Defection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models,” J. Marketing Research, vol. 43, no. 2, pp. 204-211, 2006.
- [15] W. Verbeke, D. Martens, C. Mues, and B. Baesens, “Building Comprehensible Customer Churn Prediction Models with Advanced Rule Induction Techniques,” Expert Systems with Applications, vol. 38, pp. 2354-2364, 2011.
- [16] Anuj Sharma, Dr. Prabin Kumar Panigrahi, “A Neural Network based Approach for Predicting Customer Churn in Cellular Network Services”, International Journal of Computer Applications (0975 – 8887) Volume 27– No.11, August 2011.
- [17] Wouter Verbeke, Karel Dejaeger, David Martens, Joon Hur, art Baesens, “New insights into churn prediction in the telecommunication sector: A profit driven data mining approach”, European Journal of Operation Research, 218, 2012, 211-229.
- [18] Vijaya M S “Fast single shot multiclass proximal Support Vector Machines and perceptrons” In proceedings of International conference on computing: Theory and Applications. IEEE Computer Society. 2007.
- [19] Vijaya M S “Binary classification using proximal support vector machine and lagrangian support vector machine – a comparative study”. Intelligent optimization and modelling, Alias publishing ltd. Page.no.192 to 198, 2006. ISBN 81-8424-104-6.