International Journal of Scientific Research in Computer Science, Engineering and Information Technology

© 2017 IJSRCSEIT | Volume 2 | Issue 1 | ISSN : 2456-3307

An Optimal Churn Prediction Model using Support Vector Machine with Adaboost

A. Saran Kumar, Dr. D. Chandrakala

Kumaraguru College of Technology, Coimbatore, Tamil Nadu, India

ABSTRACT

Customer churn is a common measure of lost customers. By minimizing churn, a company can maximize its profits. Companies have recognized that existing customers are most valuable assets. Customer retention is important for a good marketing and a customer relationship management strategy. In this paper, a detailed scheme is worked out to convert raw customer data into meaningful and useful data that suits modelling buying behaviour and in turn to convert this meaningful data into knowledge for which predictive data mining techniques are adopted. In this work, a boosted version of SVM which is a combination of SVM with Adaboost is used for increasing the accuracy of generated rules. Boosted versions have high accuracy and performance than non-boosted versions. The aim of churn prediction model is to detect the customers with high tendency to leave the firm and also increase the revenue for the firm.

Keywords : Churn, Adaboost, SVM, Classification and Prediction.

I. INTRODUCTION

Data classification is the process of sorting and categorizing data into various types or any other distinct class. Data classification enables the separation and classification of data according to data set requirements for various business objectives [1]. It is mainly a data management process. Classification in data mining consists of predicting a particular outcome based on the given input [2]. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome, usually called goal or prediction attribute [3]. The algorithm tries to find relationships between the attributes that would make it possible to predict the outcome. Next the algorithm is given a data set not seen before, which is called as prediction set, that contains the same set of attributes, except for the prediction attribute – not yet known. The algorithm analyses the input and produces a prediction. The prediction accuracy defines how "good" the algorithm is [4]. Classification models predict categorical class labels; and prediction models predict continuous valued functions [5]. For example, we can build a model to classify bank loan applications as either safe or risky, or a prediction model to predict the expenditures in dollars of potential customers on computer equipment given their income and occupation.

Data mining tools predict patterns and behaviors, allowing businesses to effect knowledge-driven decisions [6]. The automated, prospective analyses offered by data mining tools move beyond the analysis of past events provided by retrospective tools typical of decision support system.

Customer churn is a marketing-related term which means customers defect to another supplier or purchase less [7]. As existing customers are an important source of business profits, being able to identify customers who show signs that they are about to leave can create more income for businesses. This is especially more important for online customers, as the phenomenon of customer churn appears to be very rapid and difficult to grasp. If companies do not take measures to hold customers before their status deteriorates, the customers may never come back, resulting in wasted investment and loss of income. A timely retention strategy can keep customers, and it is the best way to retain customers [8].

II. METHODS AND MATERIAL

A. Related Work

In [9] authors study, a framework with ensemble techniques is presented for customer churn prediction directly using longitudinal behavioral data. A novel approach called the hierarchical multiple kernel support vector machine (HMK-SVM) is formulated. A three phase training algorithm for the H-MK-SVM is developed, implemented and tested. H-MK-SVM constructs a classification function by estimating the coefficients of static and longitudinal behavioral variables in the training process without transformation of the longitudinal behavioral data. The training process of the H-MK-SVM is a feature selection and time subsequence selection process because the sparse non-zero coefficients correspond to the variables selected.

In [10] author considered the two hybrid models by combining two different neural network techniques for churn prediction, which are back-propagation artificial neural networks (ANN) and self-organizing maps (SOM). The two hybrid models are, ANN combined with ANN (ANN + ANN) and SOM combined with ANN (SOM + ANN). In particular, the first technique of the two hybrid models performs the data reduction task by filtering out unrepresentative training data. The outputs as representative data are then used to create the prediction model based on the second technique. To evaluate the performance of these models, three different kinds of testing sets are considered.

In [11] authors presented an important process of developing MOD customer churn prediction models by data mining techniques. The process consists of pre-processing stage for selecting important variables by association rules that have not been applied before the model construction stage by neural networks (NN) and decision trees (DT) and four evaluation measures including accuracy, precision, recall, and F-measure all of which have not been used to examine the model performance. The source data is based on one telecommunication company providing MOD services in Taiwan, and the experimental results showed that using association rules allows the DT and NN models to provide better prediction performances over a chosen validation dataset. In particular, the DT model performs well when compared with NN model. Moreover, some useful and important rules in the DT model showed the various factors affecting a high proportion of customer churn.

In [12] authors proposed a new set of features with three new input window techniques. The new features are demographic profiles, account information, grant information, Henley segmentation aggregated call details, line information, service orders and bill and payment history. The basic idea of the three input window techniques is to make the position order of some monthly aggregated call-detail features from the previous months in the combined feature set for testing as well as for training phase. For evaluating these new features and window techniques, the two most common modeling techniques namely decision trees and multilayer perception neural networks and one of the most promising approach like support vector machines are selected as predictors.

In [13] authors presented a customer churn prediction models to detect customers with a high propensity to attrite. Prediction accuracy, comprehensibility and justifiability are three key aspects of a churn prediction model. An accurate model allow to correctly targets future churners in a retention marketing campaign, while a comprehensible and intuitive rule-set allows identifying the main drivers for customers to churn and to develop an effective retention strategy in accordance with domain knowledge. The authors provided an extended overview of the literature on the use of data mining in customer churn prediction modeling which showed that only limited attention has been paid to the comprehensibility and the intuitiveness of churn prediction models. Hence, two novel data mining techniques are applied to churn prediction modeling and benchmarked to traditional rule induction techniques such as C4.5 and RIPPER. Both Ant-Miner+ and ALBA induced an accurate as well as comprehensible classification rule-sets. Ant-Miner+ is a high performing data mining technique based on the principles of Ant Colony Optimization that allows to include domain knowledge by imposing monotonicity constraints on the final rule-set.

B. Research Methodology

The proposed architecture accepts the Customer banking parameters as input which contains the MATLAB simulation where the novel prediction algorithm is applied in bank dataset. This overall architecture in figure1 follows a Churn prediction framework from the start to end state. The users initialize the features and classes as initial parameters in which the proof process is to be evaluated.

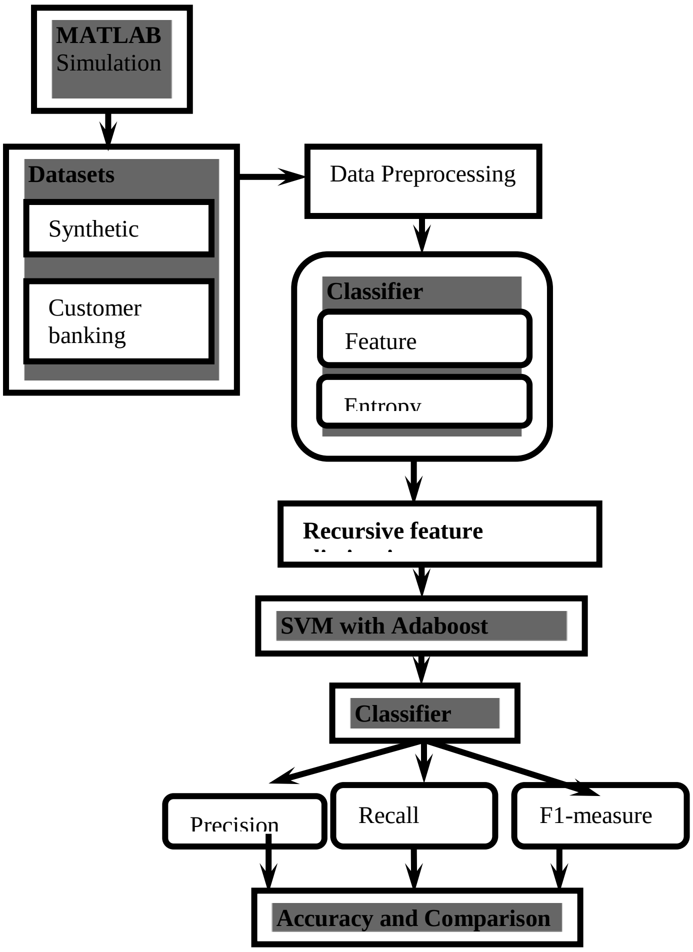

Figure 1 : Architecture of Proposed System

A. Data Preprocessing

Data preprocessing is type of data cleaning technique which plays a very important role in data classification techniques and functions. It is the first step in the adaptive relevance feature discovery mining process. In this process, there are three key steps of procedures namely Training set extraction, Feature Attribute selection and filtering methods.

The data preprocessing of untrained raw dataset is first partitioned into three groups: (1) a predetermined set of instance initiation, (2) the group of attributes (features, variables) and (3) the class of attribute. For each groups in the dataset, a reduction decision classification is constructed. For each reduction system is consequently divided into two parts: the training dataset and the testing dataset. Each training dataset uses the corresponding input features and fall into two classes: normal (+1) and abnormal (−1).

The training set feature set process to compute the cross validation classification error for a large number of features and find a relatively stable range of small error. The feature range is denoted by Ω. The optimal number of features (denoted as n*) of the training set is determined with in Ω. The complete process includes three steps:

- The Training feature selection is to select n (a preset large number) sequential features from the input X. This leads to n sequential feature sets $F_1 \subset F_2 \subset \ldots \subset F_{n-1} \subset F_n$.
- The n sequential feature sets $F_{1,}$ …, F_k, …, F_n,($1 \leq k \leq n$) to find the range of k, called Ω, within which the respective (cross-validation classification) error e_k is consistently small (i.e., has both small mean and small variance).
- Within Ω, find the smallest classification error e_k = min e_k. The optimal size of the candidate feature set, n*, is chosen as the smallest k that corresponds to e*.

The Feature attribute selection is a statistical technique that can reduce the dimensionality of data as a by-product of transforming the original attribute space. Transformed attributes are formed by first computing the covariance matrix of the original features, and then extracting its sorting. The attribute selection defines a linear transformation from the original attribute space to a new space in which attributes are uncorrelated.

B. Support Vector Machine:

SVM is a supervised learning technique from the field of machine learning applicable to both classification and regression. SVM learning machine seeks for an optimal separating hyper-plane, where the margin is maximal as shown in figure2.

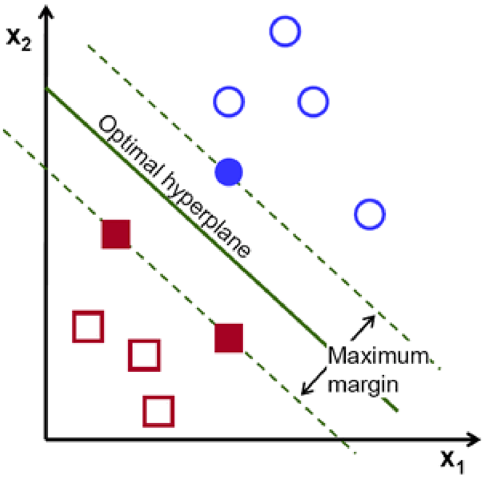

Figure 2: SVM

An important and unique feature of this approach is that the solution is based only on those data points, which are at the margin. The SVM takes an input vector $x \in R^L$, which is transformed by a predetermined feature extraction function $\phi(x) = R^N$. This feature vector is classified into one of the two classes by a linear classifier $y = sign(\langle \omega, \phi(x) \rangle + b)$ $y \in \{-1,1\}$ where $\omega \in R^N$ and $b \in R$ are the weight vector and the bias, determining placement of discrimination in the feature space respectively. The initialization of parameters *ω* and *b* is based on a supervised learning using a training set of l input-output pairs $\left\{ \left(x_i, y_i\right) \in R^l \times \{-1,1\} \right\}_{i=1}^{l}$. Assuming that the training set mapped to feature space $\left\{ \phi\left(x_i\right) \right\}_{i=1}^{l}$ is linearly separable, weight vector *ω* will be determined to maximize the margin between these two classes. Accordingly, the actual calculations of margin maximization will solve the optimizing problem with inequality constraints to

$$\text{minimize: } \frac{1}{2}\|\omega\|^2$$

$$\text{subject to : } \left\{ y_i\left(\left\langle \omega_i, \phi(x_i) \right\rangle + b\right) \geq 1 \right\}_{i=1}^{l}$$

equ.(1)

When the training set is not linearly separable in its feature space, constraints in the same optimization problem can be relaxed to

$$\text{minimize: } \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{l} \xi_i$$

subject to:

$$\left\{ y_i\left(\left\langle \omega_i, \phi\left(x_i\right) \right\rangle + b\right) \geq 1 - \xi_i \right\}_{i=1}^{l}$$

equ.(2)

where $\{\xi_i\}_{i=1}^{l}$ is non-negative slack variables that are related to the soft margin and *C* is the tuning parameter used to balance the margin and the training error. Constant *C* >0 relatively weighs penalty of training inputs residing within the inter-class margin. Both Equations (1) and (2) are optimization problems and can be solved by introducing the Lagrange multipliers $\{\alpha_i\}_{i=1}^{l}$. These multipliers transform the optimization problems to quadratic programming problems. This in essence amounts to minimization for variables *ω* and *b*, and maximization for multipliers $\{\alpha_i\}_{i=1}^{l}$. The Lagrange of these mentioned parameters is given in the Equation (3).

$$L(\omega, b, \alpha) = \frac{1}{2}\|\omega\|^2 - \sum \alpha_i \left\{ y_i\left(\left\langle \omega_i, \phi\left(x_i\right) \right\rangle + b\right) - 1 \right\}$$

equ.(3)

The optimal hyper plane is obtained using the Equation (4).

$$u(x) = \sum_{i=1}^{l} y_i \alpha_i \left\langle \phi\left(x_i\right), \phi(x) \right\rangle + b$$

equ.(4)

Typically, this optimal solution of original constrained problem, positions at a point holding for a very small fraction of *l* inequality conditions. Besides, parameter b can be obtained from (equality) constraints for one of the support vectors. The basic idea is to map *x* by nonlinearly mapping $\phi(x)$ to a much higher dimensional space in which the optimal hyper plane is found. The nonlinear mapping can be implicitly defined by introducing the kernel function *K(x,y)* which computes the inner product of vectors $\phi(x)$ and $\phi(y)$. This inner product $K(x,y) = \langle \phi(x), \phi(y) \rangle$ is the kernel function. The common kernels including, Linear Kernel Function, Gaussian Basis Function and Radial Basis Function are used in this work. In these functions *x* is the attribute value and *y* is the label value. The Gaussian basis function is defined as

$$K(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2} \right)$$

equ.(5)

where σ is standard deviation. The Linear kernel function is a type of polynomial kernel function that is defined as

$$K(x, y) = (x^T y + 1)^d$$

equ.(6)

where *d* takes the value 1, since the function is linear. The Radial basis kernel function is defined as,

$$K(x, y) = -e^{-\gamma \| x - y \|^2}$$

equ.(7)

It is very important to understand that, only the best kernel function yields the minimum error and the highest classification accuracy. The parameters used in SVM algorithm are summarized in Table 1.

Table 1 : Parameters used in SVM Classifier

Symbol	Parameters
γ	Kernel Parameter
C	Cost factor of training inputs, takes the value more than 0
ω	Weight Vector based on training set of input, output pair
b	Bias value less than 2
α	Learning Parameter
u	Optimum hyper plane
ξ	Non-negative slack parameter
$\phi(x)$	Predetermined feature extraction function
l	Number of <input, output> pair
d	Number of dimensions
σ	Standard deviation

C. SVM with Adaboost Classifier Construction

Support Vector Machine Recursive Feature Elimination (SVM-RFE) is ranking churn from bank dataset for classification. It is now being widely used for feature selection and several improvements have been recently suggested. SVM-RFE starting with all the features, removes the unwanted feature that is least significant for classification recursively in a backward elimination manner. The ranking score is given by the components of the weight vector w of the SVM as follows:

$$w = \sum_k a_k y_k x_k \qquad eqn.(1)$$

where $y_k \in l$ is the class label of the sample x_k and the summation is taken over all the training samples. A_k is the Lagrange multipliers involved in maximizing the margin of separation of the classes. If w_i denotes the component weight connecting to the attribute i, w_i^2 gives a measure the ranking of the feature i based on its effect on the margin of separation upon removal.

AdaBoost is adaptive in the sense that subsequent classifiers built are tweaked in favor of those instances misclassified by previous classifiers. AdaBoost takes as input a training set S = (x_1, y_1),…..,(x_m, y_m) where each instance, xi, belongs to a domain or instance space X, and each label yi belongs to a finite label space Y. Here we will only focus on the binary case when Y = {-1, +1}.

Each round, m = 1,…,M, AdaBoost calls a given weak or base learning algorithm which accepts as an input a sequence of training examples S along with a distribution or set of weights over the training example, W_t(i). Given such an input the weak learner computes a weak classifier, ht ∈ {-1, +1}. In general, h_t has the form ht : X → R. We interpret the sign of ht(x) as the predicted label to be assigned to instance x. Once the weak classifier has been received, AdaBoost chooses a parameter $a_t \in R$ that intuitively measures the importance that it assigns to h_t.

D. Predictive analysis

The adaptive relevance feature predictive discovery process considers the mutual-information-based feature selection for both supervised and unsupervised data. For discrete feature variables, the integral operation in (1) reduces to summation. In this case, computing mutual information is straightforward, because both joint and marginal probability tables can be estimated by tallying the samples of categorical variables in the data.

Given two random variables x and y, their mutual information is defined in terms of their probabilistic density functions $p(x)$, $p(y)$, and $p(x, y)$:

$$MI(x,y) = \int\int p(x,y) log \frac{p(x,y)}{p(x)p(y)} dxdy \qquad eqn.(2)$$

In Maximum Relevance discovery, the selected features x_i are required, individually, to have the largest mutual information $MI(x_i, c)$ with the target class c, reflecting the largest dependency on the target class. In terms of genetic search, the m best individual features, i.e., the top m features in the descent ordering of $MI(x_i; c)$, are often selected as the m features.

However, when at least one of variables x and y is continuous, their mutual information $MI(x; y)$ is hard to compute, because it is often difficult to compute the integral in the continuous space based on a limited number of samples. One solution is to incorporate data discretization as a preprocessing step.

Given N samples of a variable x, the approximate similarity function $Simm$ (x) has the following form:

$$Simm(x) = \frac{1}{N}\sum_{i=1}^{N} \delta(x - x^i, h) \quad eqn.(3)$$

where $\delta(.)$ is the sampling window function as explained below, $x^{(i)}$ is the i^{th} sample, and h is the window width.

III. CONCLUSION

This paper presents an enhanced method such as SVM with Adaboost Classification using Feature Discovery (FD) based prediction method which combines classifications of SVM, NBTree and SVM Adaboost to solve the problem of high dimensional classification. The proposed method accepts a hybrid approach for extracting rules from SVM for customer relationship management (CRM) with Adaboost classification. The proposed hybrid approach can be used to achieve higher classification accuracy which can be used to predict churners in order to provide personalized offers and to increase revenue for the firm.

IV. REFERENCES

[1]. V. N. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, NewYork Inc., NY, USA, 1995.

[2]. I. Guyon, J. Weston, S. Barnhill, V.N. Vapnik, Gene selection for cancer clas-sification using support vector machines, Machine Learning 46 (1–3) (2002)389–422.

[3]. K.J. Kim, Financial time series forecasting using support vector machines, Neu-rocomputing 55 (1/2) (2003) 307–319.

[4]. S. Ben-David, M. Lindenbaum, Learning distributions by their density levels:a paradigm for learning without a teacher, Journal of Computer and SystemSciences 55 (1997) 171–182.

[5]. C.-H. Wu, Y. Ken, T. Huang, Patent classification system using a new hybridgenetic algorithm support vector machine, Applied Soft Computing 10 (4)(2010) 1164–1177.

[6]. S. Chowdhury, J.K. Sing, D.K. Basu, M. Nasipuri, Face recognition by general-ized two-dimensional FLD method and multi-class support vector machines,Applied Soft Computing 11 (7) (2011) 4282–4292.

[7]. H. Azamathulla, Md. Fu-Chun Wu, Support vector machine approach to forlongitudinal dispersion coefficients in streams, Applied Soft Computing 11 (2)(2011) 2902–2905.

[8]. V.N. Vapnik, Statistical Learning Theory, John Wiley & Sons, New York, USA,1998.

[9]. Zhen-Yu Chen, Zhi-Ping Fan, Minghe Sun, "A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioral data," European Journal of Operational Research 223 (2012) 461–472.

[10]. Chih-Fong Tsai, Yu-Hsin Lu, "Customer churn prediction by hybrid neural networks," Expert Systems with Applications 36 (2009) 12547–12553

[11]. Chih-Fong Tsai, Mao-Yuan Chen, "Variable selection by association rules for customer churn prediction of multimedia on demand," Expert Systems with Applications 37 (2010) 2006–2015

[12]. B.Q. Huang, T.-M. Kechadi, B. Buckley, G. Kiernan, E. Keogh, T. Rashid, "A new feature set with new window techniques for customer churn prediction in land-line telecommunications" Expert Systems with Applications 37 (2010) 365.7–3665

[13]. Wouter Verbeke, David Martens, Christophe Mues, Bart Baesens, "Building comprehensible customer churn prediction models with advanced rule induction techniques" Expert Systems with Applications 38 (2011) 2354–2364.

[14]. P.C. Pendharkar, "Genetic algorithm based neural network approaches for predicting churn in cellular wireless network services", Expert System Application 36 (2009) 6714–6720.

[15]. Yaya Xie, Xiu Li, E.W.T. Ngai, Weiyun Ying, "Customer churn prediction using improved balanced random forests", Expert Systems with Applications 36 (2009) 5445–5449.