Customer Retention of MCDR using 3SCDM Approaches

Suban Ravichandran [1], Chandrasekaran Ramasamy [2]

Assistant Professor, Department of Computer Science and Engineering, Annamalai University, Chidambaram, India [1]

Professor, Department of Computer Science and Engineering, Annamalai University, Chidambaram, India [2]

Abstract: In telecommunication industry satisfying customers' needs plays a vital role to retain them within their network for longer duration of time. A well-known fact in the telecommunication industry is that the competition among industries is very fierce. The acquisition of new and resourceful customers has become difficult and often very expensive. Subsequently customer retention has become more and more important. Data Mining can determine characteristic customer clusters on the basis of collected historic data points from customers - such as for instance the frequency and timely distribution of customer's usage of services (calls, text messages, MMS, navigation, mail exchange). For each of these customer patterns the company can then offer tailored customer life cycle messages and offers. Implementing the Three-Stage Classifier based Data Mining (3SCDM) approach, an operator can predict churn, incentives may be offered to the customers for successful retention. The proposed system is evaluated by implementing Chi-Square (Chi^2) Feature Reduction method along with 3SSCDM approach. Combination of Bayesian Network – RBFNet – RT, Bayesian Network – RBFNet – J48 and Bayesian Network – RBFNet – MLP classifiers are used in Three-Stage Classifier (TSC). On comparing the performance based on accuracy and time taken, Bayesian Network – RBFNet – RT with Chi-Square method performs well by 89.31% and 8.03 secs respectively. This inference can be used for identifying the prospective 3G customers in the network.

Keywords: Three-Stage Classifier based Data Mining, PAKDD 2006, MCDR, Bayesian Network, RBFNet, RT, J48, MLP, Chi-Square, and WEKA.

I. INTRODUCTION

Data mining can include esteem by being beneficially connected to challenges in the telecom division. It is essential for the business to center back on enhancing the client experience. This is turning out to be progressively vital to service differentiation and retaining customers in this aggressive and monetary based environment is challenging. Remembering that the expense of retaining clients has become expensive; diminishing churn rates can offer huge funds saving to telecom industries.

Customer Retention issues [1] in a mobile market addresses to allow researchers in contributing solution for several critical problems that should be followed. First, convert the MCDR dataset from unbiased stage to suitable format for processing. Complexity of the process can be reduced by applying the Chi-Square feature selection techniques. Chi-Square Method is a statistical approach applied to sets of categorical data to evaluate the appropriate set of featured data available in the original set.

3SCDM is the combination of three classifiers, required for the betterment of the analytical based results and attain reduced error rate. In 3SCDM, the outcome of the first stage classifier is directly given as input to the second stage classifier and outcome of the second stage classifier is directly given as input to the third stage classifier. The work has been carried out by considering the combination of classifiers like Bayesian Network – RBFNet – RT,

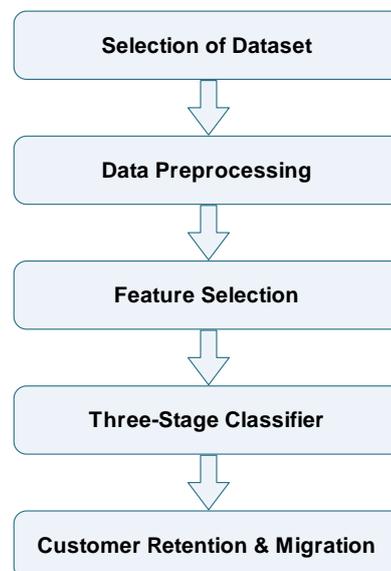Bayesian Network – RBFNet – J48 and Bayesian Network – RBFNet – MLP.

Fig 1

Fig 1. Block Diagram for MCDR using 3SCDM Approach
Data Mining (DM) is distinct as the process of extracting implicit, previously unknown, and most constructive information available in the data [2]. DM acts as a central

step for analysing data in Knowledge Discovery in Databases (KDD). KDD manages discovering designs, rules, connections, deviations, and uncommon occasions in colossal volume of information put away in relational databases and data warehouses [3]. The KDD process involves the following steps: Data Gathering, Data Cleansing, Feature Extraction, and Data Mining, Visualization of data and verification and evaluation of results [4].

Mobile operators have stored collecting massive information of users in their network based on the usage of their services provided. They have come forward to utilize those massive data for optimizing network services, strategically positioning service personnel, perform customer service requests, and analyze the behavioural aspects of customers and more. This analysis could provide information like human behaviour, communication patterns, geographical behaviour, mode of service usage, economic consuming behaviour, etc.

Concluding the Introduction, the rest of the paper is organized as described below. Section II describes an overview of Dataset and different modelling methods are discussed in Section III. Section IV provides a brief discussion about Simulation Setup, Analysis and comparison of different modelling techniques are presented in Section V and Section VI concludes the carried out research and possible future works.

II. METHODOLOGY

A. Data Preprocessing

The dataset PAKDD 2006 is available in "dat" for in default. This format is imported in Excel and converted into Comma Separated Value (CSV) format. The converted file is then fed into the Preprocessing Tool developed [5] in PHP and MySQL as shown in Fig 2. The developed tool removes the duplicate instances and attributes values available in the dataset. The tool is developed in five stages namely.

i.   Importing Data.
ii. Remove Duplicate Instances Values.
iii. Remove Duplicate Column (Attribute) and Null or Zero Values.
iv. Conversion of Text into Numeric Values.
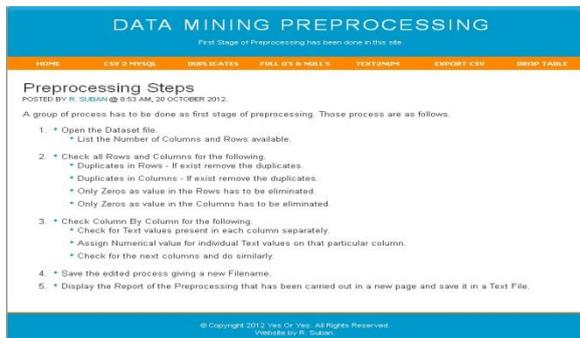v.   Export preprocessed data in CSV format.

Fig 2.   Developed Data Preprocessing Tool

Once the data is preprocessed the data is exported in the CSV format as a file and fed for further modelling. Each and individual steps ii, iii, iv are designed in such a way they are not interrelated. Those steps can be executed separately without each other. The end result of the preprocessing contains the same number of attributes and instances as original since no duplicates are available. The resultant dataset contains 251 attribute with 18000 instances.

B. Replace Missing Values

Missing value is the number (percentage) of instances in the data for which the attribute is unspecified. We use a filter called "Replace Missing Values", which replaces all missing values for all nominal and numeric attributes in a dataset.

C. Chi-Squared Feature Selection Techniques

Chi-square (χ^2) test based feature selection is one another commonly used method in feature reduction [6], [7], [8]. This chi-square computes the value of the chi-squared statistic with respect to the class to evaluate the worth of the feature set. With the assumption of initial hypothesis H_0 that the two features are unconnected and it is experienced by chi-squared formula:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \qquad \dots (1)$$

Where E_{ij} is the theoretical expected frequency and O_{ij} is the observed frequency, emphasized by the null hypothesis. Usually the value of χ^2 is proportionate to the evidence against the hypothesis H_0. On Applying Chi-Square Attribute Evaluator to the dataset containing 251 attribute and 18000 instances, 13 attributes has been selected.

TABLE 1. Selected Attribute by Chi – Square Attribute Evaluator

Method	Positions of Selected Attributes
Chi – Square Attribute Evaluator	11, 23, 26, 77, 85, 88, 89, 98, 102, 103, 132, 219, 251.

D. Classifiers

i. Bayesian Network

The Bayesian network (BN) structures [13] mainly encode variable independence. BN have high computational complexity in a learning of its structure. Variable autonomy in instruct Conditional Probability Tables (CPT) having more obligation in the execution, rather than in learning structure. A full Bayesian model is called Full Bayesian system Classifiers (FBC).

A Bayesian Network (BN) comprises of a coordinated non-cyclic diagram and a set of likelihood appropriations, where hubs and curves in non-cyclic diagram speak to

arbitrary variables and the direct connections between variables individually, and the set is the arrangement of nearby distributions for each node.

ii. Radial Basis Function Network

Radial basis function (RBF) network is a process of a three–layer feed–forward network consisting an input units, a hidden layer with the Gaussian function units and an output layer. The hidden layer has a center which is a vector in each units. The centers is defined as the process of comparing input vectors in the process of producing radially symmetrical responses. The hidden layer responses are scaled and combined by the connection weights in order to produce the network output.

In the procedure of achieving an RBF network with adaptively variable structure, it is proposed a sliding data window [11], which is a first-in-first-out sequence with fixed width. By receiving a new sample, the sliding window is updated by adding the newer data and discarding the older data. The sliding window data (W_{SD}) is stated in the form as given below:

$$W_{SD} = [(x_t - N + 1, y_t - N + 1), \dots, (x_t, y_t)] \ \dots (3)$$

Where N is the width of the sliding data window.

iii. Random Tree

Random tree models have been extensively developed in the field of Machine Learning in the recent years. With k random features at each node, a random tree is a tree drawn at random from a set of possible trees. In this context "random" means that each tree in the set of trees has an equal chance of being sampled. Another way of saying this is that the distribution of trees is "uniform". Random trees can be generated efficiently and the combination of large sets of random trees generally leads to accurate models. Random tree models have been extensively developed in the field of Machine Learning in the recent years [5].

iv. J48 Algorithm

J48 is a tree based learning approach. It is developed by Ross Quinlan which is based on iterative dichotomiser (ID3) algorithm [9]. J48 uses divide-and-conquer algorithm to split a root node into a subset of two partitions till leaf node (target node) occur in the tree. Given a set T of total instances the following steps are used to construct the tree structure.

Step 1: If all the instances in T belong to the same group class or T is having fewer instances, than the tree is leaf labeled with the most frequent class in T.

Step 2: If step 1 does not occur then select a test based on a single attribute with at least two or greater possible outcomes. Then consider this test as a root node of the tree with one branch of each outcome of the test, partition T into corresponding T1, T2, T3........, according to the result for each respective cases, and the same may be applied in recursive way to each sub node.

Step 3: Information gain and default gain ratio are ranked using two heuristic criteria by algorithm J48.

v. Multi-Layer Perceptron

Multi-Layer Perceptron (MLP) [9] is a feed forward neural network with one or more layers between input and output layer. Feed forward means that data flows in one direction from input to output layer (forward). This type of network is trained with the backpropagation learning algorithm. MLPs are widely used for pattern classification, recognition, prediction and approximation. Multi-Layer Perceptron can solve problems which are not linearly separable.
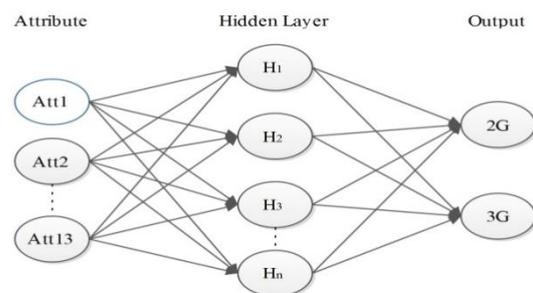

Fig 3.   Layer Structure for MLP

The Multi-Layer Perceptron is an example of an artificial neural network that is used extensively for the solution of a number of different problems, including pattern recognition and interpolation. It is a development of the Perceptron neural network model, that was originally developed in the early 1960s but found to have serious limitations.

III. SIMULATION SETUP

A. Dataset

A dataset is roughly equivalent to a two-dimensional spreadsheet or database table. In this paper we discussed a dataset which was successfully launched a 3G network by an Asian Telco operator (A Mobile Network operator). Being provided with a dataset containing a sample set of 20,000 2G and 4,000 3G network customer with 251 attributes, the operator wants to identify the willingness to switch over to 3G from 2G network using existing usage and demographic data [12]. The training dataset consists of three quarters with 15K 2G and 3K 3G customer data. The remaining one quarter of the dataset 5K 2G and 1K 3G is used as Test dataset for prediction process [13].

Dataset utilized here is PAKDD 2006. The PAKDD information mining rivalry introduces an issue on the surely understood use of credit scoring. The offer of credit for potential customers is an essential administration for empowering utilization in the business sector. A unique example dataset of 20,000 2G system clients and 4,000 3G system clients has been furnished with more than 200 information fields. The objective all out variable is "Customer_Type" (2G/3G) [5].

B. Proposed Work

The result of the preprocessed PAKDD 2006 dataset is used in chi - square attribute evaluator method and the result is an input to a three-stage classifier. Three-Stage Classifier has the combinations of Bayesian Network with RBFNet with RT or J48 or MLP as shown in the Fig 3.

Three-Stage Classifier has a process of Bayesian Network as Stage – I, RBFNet as Stage – II and Random Tree or J48 or MLP as Stage – III. The resultant confusion matrix obtained is used for the calculation of performance metrics and Churn or Customer Retention.
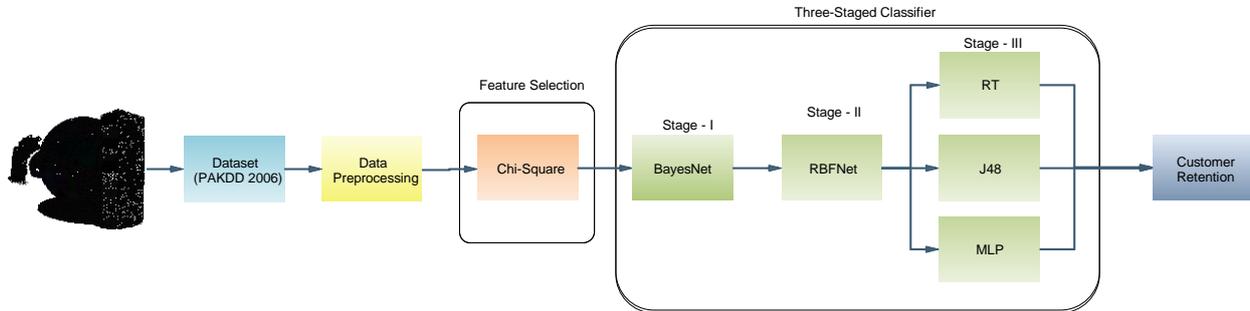

Fig 4.   Work Flow Diagram for MCDR using TSCDM Approach

IV. EXPERIMENTAL RESULTS

The proposed approaches have been experimented with the desired dataset PAKDD 2006 in order to exactly locate the performances of the system. Performance measures such as Time Taken, Accuracy, Sensitivity and Specificity are measured and considered for the evaluation of the proposed approaches.
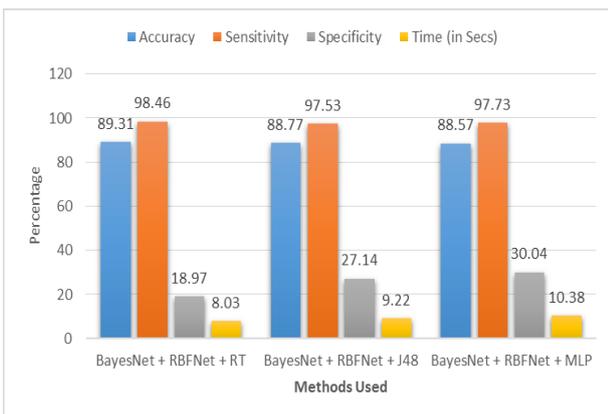

Fig 5.   Performance Measures of Three-Staged Classifier with Chi^2

The churn has been estimated with the greater exactitude for all the three-staged classifiers combinations with chi^2 methods. The proposed methods are ranked based on the estimation of churn listed in the Table II.

TABLE 2.   Churn Prediction and Rank based on Churn

Algorithm	Churn	Rank
BNet - RBFNet - RT	2233	1
BNet - RBFNet – J48	2210	2
BNet - RBFNet - MLP	2156	3

V. CONCLUSION AND FUTURE SCOPE

The inevitable stab tackled investigating the practice of Data Mining approaches in the field of Mobile Communication through the customer retention modelling system. On the basis of untried evidence, the new proposal of customer retention model is realized and their performance is precisely estimated. The three-staged classifiers with chi^2 are extensively providing fulfilling outcome in the analysis of 2G dataset PAKDD 2006. The proposed BNet - RBFNet - RT model outperforms all the other classification algorithms of accuracy and with higher dimensional reduction leads to less processing time about 89.31% and 8.03 secs respectively. As for as the churn concern, the BNet - RBFNet - RT model scores the peak. In further, the proposed work can be carried out based on the perspective of clustering methods for considering the behavioural analysis of customers and classification of 2G and 3G.

REFERENCES

[1] Polo, Yolanda, F. Javier Sese, and Peter C. Verhoef. "The effect of pricing and advertising on customer retention in a liberalizing market." Journal of Interactive Marketing 25.4 (2011): 201-214.

[2] J. Han And M. Kamber, "Data Mining: Concepts and Techniques", San Francisco, Morgan Kauffmann Publishers, 2001.

[3] Dzeroski, Saso, and Nada Lavrač, eds. Relational data mining. Springer, 2001.

[4] Hafez, Alaaeldin M. "Knowledge Discovery in Databases." (2008).

[5] Suban Ravichandran, Vijay Bhanu Srinivasan and Chandrasekaran Ramasamy, "Comparative Study on Decision Tree Techniques for Mobile Call Detail Record", Journal of Communication and Computer, Volume 9, Number 12, pp.1331-1335, December 2012.

[6] Novaković, Jasmina, Perica ŠTRBAC, and Dušan Bulatović. "Toward optimal feature selection using ranking methods and classification algorithms." The Yugoslav Journal of Operations Research 21, no. 1 (2011).

[7] Muntean, Cristina Ioana, Darie Moldovan, and Ovidiu Veres. "A personalized classification of employment offers using data mining methods." International Journal of Mathematical Models and Methods in Applied Science 3, no. 5 (2011): 525-532.

[8] Balachandran, K., and R. Anitha. "Dimensionality reduction based on the classifier models: Performance Issues in the prediction of Lung cancer." In Software Engineering (CONSEG), 2012 CSI Sixth International Conference on, pp. 1-4. IEEE, 2012.

[9] Suban Ravichandran, Vijay Bhanu Srinivasan and Chandrasekaran Ramasamy, "Measuring Accuracy of Classification Algorithms for Chi-Square Attribute Evaluator in MCDR", "International Conference on Advances in Engineering And Technology Convergence, At Bangkok, Thailand", pp. 6 – 9, April 28th, 2013.

[10] Wong, Man Leung, and Kwong Sak Leung. "An efficient data mining method for learning Bayesian networks using an evolutionary algorithm-based hybrid approach." IEEE Transactions on Evolutionary Computation 8.4 (2004): 378-404.

[11] Yin, Jian-Chuan, Li-Dong Wang, and Ni-Ni Wang. "A variable-structure gradient RBF network with its application to predictive ship motion control." Asian Journal of Control 14.3 (2012): 716-725.

[12] The 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2006), http://www3.ntu.edu.sg/SCE/pakdd2006/competition/overview.htm.

[13] Dr. Tay Joc Cing, Nanyang, Dr. Andrew Watkins, Hanny Yulius Limanto, "Classification of Mobile Phone Customers", PAKDD 2006.

BIOGRAPHIES

Suban Ravichandran received B.E. Degree in Information Technology from Bharathidasan University in 2003 and M.E. Degree in Computer Science and Engineering from Annamalai University, Tamil Nadu, India in 2007. He is currently working as an Assistant Professor at the Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar, Tamil Nadu, India. His research interests include Data Mining, Computer Networks and Mobile Computing.

Mr. R. Suban is Life Member of the Computer Society of India, Indian Society for Technical Education, International Association of Computer Science and Information Technology and International Association of Engineers. He is a Global Member in Internet Society. He is a Research Volunteer in International Journal of Advances in Engineering & Technology. And he is also a Student Life Member in American Finance Association.

Chandrasekaran Ramasamy received the B.E Degree in Electrical and Electronics Engineering from Madurai Kamaraj University in 1982 and the MBA (Systems) in 1995 from Annamalai University, M.E in Computer Science and Engineering from Anna University and Ph.D. Degree in Computer Science and Engineering from Annamalai University, Tamil Nadu, India in 1998 and 2006, respectively. He is currently working as a Professor at the Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar, Tamil Nadu, India. From 1999 to 2001 he worked as a software consultant in Etiam, Inc., California, USA. He was formerly Registrar, Anna University, Tiruchirappalli during the period 2007 – 2010.

Dr. RM. Chandrasekaran has conducted Workshops and Conferences in the Areas of Multimedia, Business Intelligence and Analysis of algorithms, Data Mining. He has presented and published more than 45 papers in conferences and journals and is the author of the book Numerical Methods with C++ Program (PHI, 2005). His Research interests include Data Mining, Algorithms, Networks, Software Engineering, Network Security, and Text Mining. He is Life member of Computer Society of India, Indian Society for Technical Education, Institute of Engineers and Indian Science Congress Association.