Scientific Journal of Impact Factor (SJIF): 4.14

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 3, Issue 12, December -2016

STUDY ON CHURN PREDICTION IN TELECOM INDUSTRY

[1]N.Kamalraj and [2]Dr.A.Malathi

[1]*Assistant Professor and* [2]*Assistant Professor*
[1]*Department of Information Technology, Dr.SNS Rajalakshmi College of Arts & Science, Coimbatore, TN, India.*
[2]*PG and Research Department of Computer Science, Govt. Arts College, Coimbatore, TN, India.*

Abstract - *Prediction of customers who are leaving a company is called as churn prediction in telecommunication. Churns are classified into two types. They are: voluntary and involuntary. Involuntary churners are identified and classified the customer who eliminates the subscribers list. Voluntary type of churners is difficult to resolve the customer who stops the service provider. Voluntary type of churners happens on the telecom industry. Consequently, the predictive model is concerned on predicting customers and analyzing the behavior which causes the problems. Predicting the customers who are expected to churn is the key objective carried out using the data mining techniques. Churners take place on many business areas, so the key aim of the work is to develop a useful tool with the data mining techniques for predicting the customer churn risk. The research work is concentrated on predicting the model with the results and infers the feasible reasons on churn in telecom industry.*

Keywords: *Churn prediction process, Churners, Telecom industry, involuntary churners, Voluntary churners*

1. INTRODUCTION

Data mining methods lie at the intersection of artificial intelligence, machine learning, statistics and database systems. Data mining techniques helps in building the prediction models to discover future developments and actions allowing the organizations to take smart decisions derived from the knowledge from data.

Churn prediction is an application of consumer performance in data mining. Churn is a key issue faced through an enterprise and denoted the cost of extending a new customer is nearly five times higher than the cost of maintaining an old customer. Because the competitiveness of the enterprise market is declined through churn, churn prediction is carried out through data mining to enhance the customer maintenance. Companies identify the consumers who are not unenthusiastic to move near a competitor through churn prediction. After that, suitable advertising operations are used to preserve and hold the customers. Churn prediction permits companies to enhance the efficiency of customer retention operations and to minimize the costs linked with churn.

Churn prediction process on telecom industry is an essential research area for the recognition of the faults. Customer churn is characterized as the loss of customers as they leave to their competitors. It is key issue in telecom industries as taking new customer's costs five to six times higher than maintaining the existing ones. In telecommunication industry, the churn is also called as customer attrition or subscriber churning. It is termed as the phenomenon of loss of a customer. It is determined by the rate of churn and significant indicator for organizations. The process of movement from one provider to another is happening because of the better rates or services or different advantages which the competitor company provides while signing up.

In the business environment, churn indicates customer migration and loss of value. Churn rate is calculated as the percentage of customers who end connection with the organization or with customers receiving their services. In new associations, there are large demand that predicts the customers to maintain them promptly by reducing the costs and risks. It also increases the efficiency and competitiveness. They are used in market advanced analytics tools and applications designed

to identify the large amount of data inside the groups. It also creates prediction derived from the information attained by examining and exploring the data.

This paper is organized as follows: Section II discusses reviews on churn prediction in telecommunication industry, Section III describes the existing churn prediction techniques in telecom industry, Section IV identifies the possible comparison between them, Section V explains the limitations as well as the related work and Section VI concludes the paper, key areas of research is given to develop a useful tool with the data mining techniques for predicting the customer churn risk.

2. LITERATURE REVIEW

Existing Privacy-Preserving Data Mining (PPDM) is designed by Fosca Giannotti., et al., (2013) encrypts its client data with (E/D) module. E/D module regains the true returned patterns with efficient plan of incremental synopsis. PPDM ensure the each transformed item and contain the churn by identifying the background knowledge. However the issues (i.e., churn) occurs on this system is unable to cluster k-privacy items in group for easy identification of the problem set. To implement with the alternative clustering, Fast Clustering-Based Feature Subset Selection (FAST) is introduced by Qinbao Song., et al., (2013). Features are divided into clusters using graph-theoretic clustering methods for creating a subset of useful and independent features. But, the recognized properties of feature space clustering with the fault set are not assured.

Classification Rules based on the time series analysis is presented by Dominik Fisch., (2011) performs the segmentation piecewise with the polynomial modeling. This polynomial type of modeling is extremely fast processing step in only one pass over the time series. Classification Rule construct the dynamic classifier with the static one and also maintains the different univariate Gaussian set but comprehensibility rule set is not addressed on predicting the churns (i.e., faults). Distributed Strategies for outlier mining using classical nested-loop approach designed by Fabrizio Angiulli., et al., (2013) to determine the number of local active objects (i.e., faults). Outlier detection solves set of top-n distance but rule based on the data mining procedure is not obtained with good solutions.

Regularization-based algorithm called Ranking Adaptation SVM (RA-SVM) was briefed by Bo Geng., et al., (2012) with an effective domain-specific ranking model to perform the prediction. Ranking Support Vector Machines (Ranking SVM) is one of the most effective learning to rank algorithms but fail to quantitatively measure specific target domain results. Fast Distributed Mining (FDM) with association rules designed by Tamir Tassa., (2014) calculates the union of subsets to identify the player's privacy level. Problem arises on subgroup discovery with data mining rule on horizontally partitioned faulty data.

3. DECISION TREE BASED CHURN PREDICTION IN TELECOM INDUSTRY

Churn customer is denoted as the customer one who disconnects the connection with existing company and joins as a customer of another competitor company. The management assumed to find out the customer turnover is named as churn management. Churn is described as a withdrawing the connection with a company. Reducing churn is essential as attaining new customer's costs higher than preserving the existing customers. Churn rate calculates the number of customers leaving a company in a particular interval of time. It is utilized in business displaying the number of customers leaves the company or stay in. The churn rate is also employed in telecommunication industry. Customer movement starting with one supplier then onto the next in media transmission industry is called customer churn and the administrator's procedure to keep up the useful customers calculated as churn association.

Decision tree is employed to predict future trends and to remove the models depending on the interrelated decisions. It works on the key of arranging information into a specific class with their properties. Inward nodes take after the root node by covering all event potential outcomes. In this manner, a tree is shaped with its one of a kind curve depicting specific reactions.

3.1 Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data

The key aim of choosing the subset of features with respect to the target ideas, feature subset selection is an effective method for minimizing the dimensionality, eliminating the inappropriate data, increasing learning accuracy and enhancing the comprehensibility. Many feature subset selection methods are designed for machine learning functions. The methods are classified into four approaches: the Embedded, Wrapper, Filter, and Hybrid approaches.

Fast clustering bAsed feature Selection algoriThm (FAST) functions in two steps. In former step, features are separated into clusters using graph-theoretic clustering methods. Later, the representative feature is connected with the target classes chosen each cluster to create the final subset of features. Features in various clusters are independent, the clustering based strategy of FAST contains a high probability of creating subset of useful and independent features. The feature subset selection algorithm FAST analyzed the available image, microarray and text data sets.

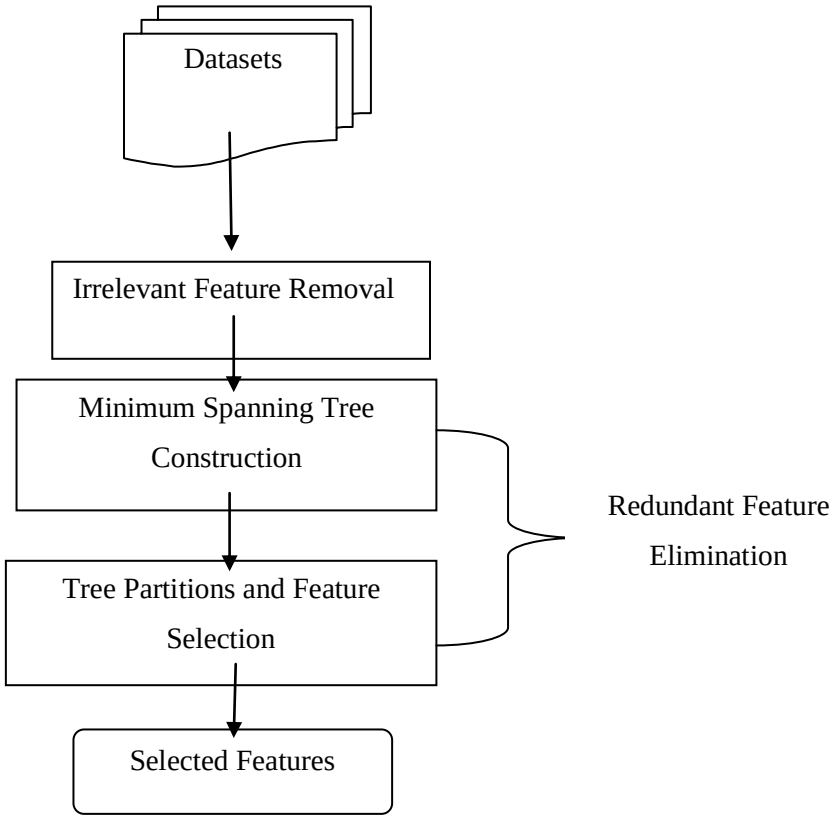

Figure 1 Feature subset selection algorithm

Figure 1 explains the feature subset selection algorithm. At first, the components are gathered from the data sets. In the wake of gathering the components, the unessential features are expelled. Redundant feature elimination process is carried out after removing the features using minimum spanning tree construction. Finally, the features get selected.

3.2 Secure Mining of Association Rules in Horizontally Distributed Databases

An alternative protocol is designed called secure mining protocol for the secure calculation of the combination of private subsets. The secure mining protocol improves in terms of simplicity, efficiency and privacy. The protocol fails to rely on commutative encryption and oblivious transfer. When the solution is not secured, it sends out excess information to a small number of feasible associations that reveals information to single players. The additional information is not more

responsive than the information revealed. The protocol computes a parameterized family of functions called threshold functions where two extreme cases matched to the issues of computing the union and intersection of private subsets.

Temporal data mining addresses tasks like segmentation, classification, clustering, forecasting and indexing of time series, event sequences, or sections of time series or sequences for more information on temporal data mining. Applications manage and control the financial, biological, medical, meteorological, or technical time series or sequences. A new approach is designed to determine and apply classification rules for time series, segments of time series or sequences of segments. It is revealed how a classifier derived from classification rules is employed for pattern discovery (i.e.,) the recognition of approximately repeated subsequences of time series. An approach is designed to extract the logical classification rules for time series analysis called SwiftRule. In addition it explains how a time series are segmented and modeled in a suitable form. The similarity of time series or segments are calculated derived from the models and rule sets (i.e., classifiers) of the form sketched are found and applied to categorize segments or sequences of segments.

3.3 Ranking Model Adaptation for Domain-Specific Search

Learning to rank is a type of learning-based information retrieval techniques, specialized in learning a ranking model with documents labeled with their relevancies to some queries. The model ranks the documents revisited to an arbitrary new query mechanically. Derived from various machine learning methods, e.g., Ranking SVM, Rank Boost, Rank Net List Net, Lambda Rank the learning to rank algorithms with the performances in information retrieval, particularly web search.

An adaptation of ranking models is designed by developing the labeled data from auxiliary fields directly unreachable because of privacy problem or data missing. Model adaptation is attractive than data adaptation, as the learning complexity is connected with the size of the target domain training set which is smaller than the size of auxiliary data set. The three issues of ranking model adaptation are given by: adapt ranking models discovered for broad-based search or verticals or a new domain consequently the amount of labeled data in the target domain is minimized as the performance need is guaranteed. In addition, it also changes the ranking model efficiently. It also uses domain-specific features to enhance the model adaptation.

The first problem is clarified with the ranking adaptability measure that quantitatively computes whether a current ranking model is conformed to the new domain and figures the potential results for the adjustment. The second problem is addressed by planning the regularization framework and a ranking adaptation SVM (RA-SVM) algorithm. The algorithm is a black box ranking model adaptation that demands for predictions from the ranking model than the internal illustration of the model or the data from the auxiliary areas. With black-box adaptation property, the flexibility and the efficiency are achieved. To determine the third issue, documents that are equivalent as domain-specific feature space include consistent rankings (i.e., images that are similar in their visual feature space should be ranked into similar positions).

3.3.1 Top-k Unexplained Sequences in Time-Stamped Observation Data

The main aim of the technique is to find an unexplained sequence detector. It is used to identify subsequences of the observation data called unexplained sequences where the recognized models are failed to explain with a confidence. In addition, unexplained sequences are not collected by the existing activity models in set activities. Each unexplained sequence is revealed to a domain expert who adds an observed sequences or generalizations to the known list of good or bad activities. Unexplained sequences permit an application to recognize behavior that are not recognized or imagined by experts and to incorporate them to rising body of knowledge.

4. PERFORMANCE ANALYSIS OF CHURN PREDICTION

In order to compare the churn prediction using different techniques, number of subscribers is taken to perform the experiment. Various parameters are used for churn prediction with fuzzy rule set on telecom industry.

4.1 Accuracy

Accuracy is defined as the ratio of number of correct predictions to the total number of predictions. It is measured in terms of percentage (%).

$$Accuracy = \frac{Number\ of\ exact\ predictions}{Total\ number\ of\ predictions}$$

Table 4.1 Tabulation of Accuracy for Churn Prediction Techniques

Number of Subscribers	Accuracy (%)			
	FAST	Secure Mining Protocol	SwiftRule	RA-SVM
10	56	62	59	65
20	58	66	63	68
30	62	69	66	72
40	65	73	69	75
50	69	76	72	78
60	73	79	75	81
70	76	82	79	84

Accuracy comparison takes place on existing Fast clustering bAsed feature Selection algoriThm (FAST), Secure Mining Protocol**,** SwiftRule and Ranking Adaptation SVM (RA-SVM).

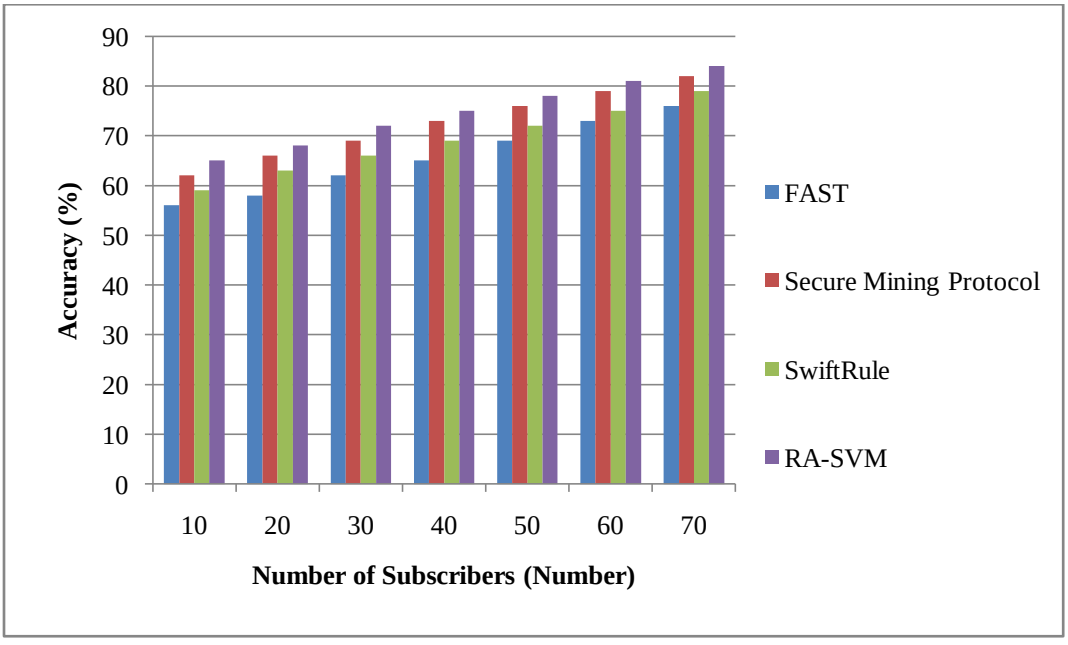

Figure 4.1 Accuracy for Churn Prediction Techniques

From figure 4.1, accuracy for churn prediction rate is evaluated. Accuracy of Ranking Adaptation SVM (RA-SVM) is higher than Fast clustering bAsed feature Selection algoriThm (FAST), Secure Mining Protocol and SwiftRule. Research in Ranking Adaptation SVM (RA-SVM) has 12.38% higher accuracy than Fast clustering bAsed feature Selection algoriThm (FAST), 3.11% higher accuracy than Secure Mining Protocol and 7.71 % higher accuracy than SwiftRule.

4.2 Clustering Efficiency

Clustering Efficiency is defined as the ratio of amount of cluster formed to the maximum cluster forming data size. It is measured in terms of percentage (%).

$$Clustering\ Efficiency = \frac{Amount\ of\ Cluster\ formed}{Maximum\ Cluster\ forming\ data\ sizes}$$

Table 4.2 Tabulation of Clustering Efficiency for Churn Prediction Techniques

Number of Subscriber (Number)	Clustering Efficiency (%)			
	FAST	Secure Mining Protocol	SwiftRule	RA-SVM
10	56	62	69	59
20	59	65	73	63
30	62	69	76	65
40	65	73	79	69
50	68	76	82	71
60	72	79	85	74
70	75	83	89	78

Clustering Efficiency comparison takes place on existing Fast clustering bAsed feature Selection algoriThm (FAST), Secure Mining Protocol, SwiftRule and Ranking Adaptation SVM (RA-SVM).

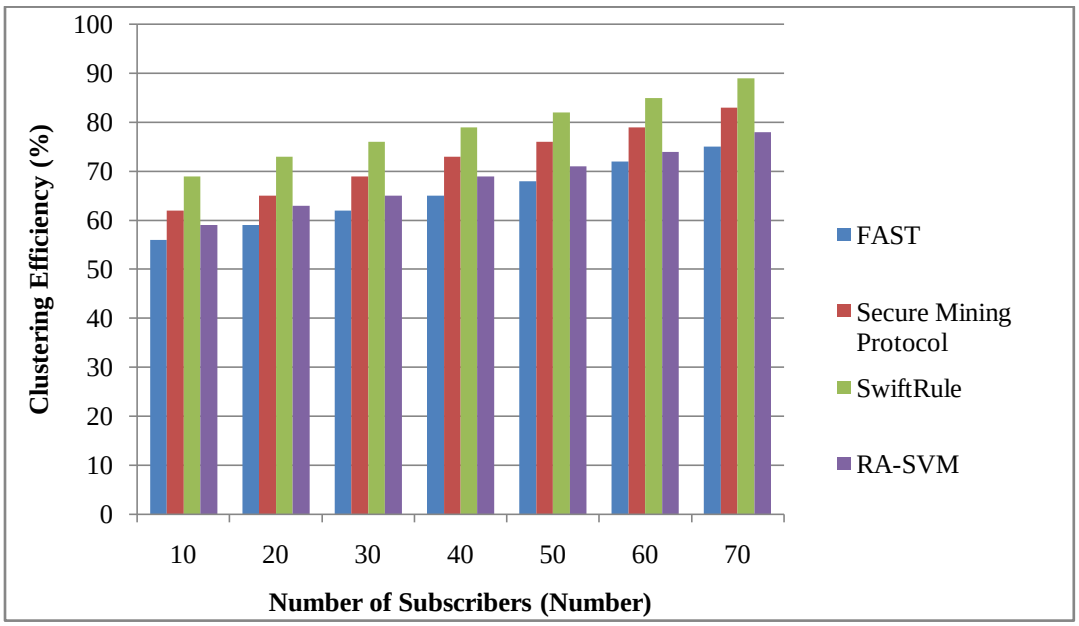

Figure 4.2 Clustering Efficiency for Churn Prediction Techniques

From figure 4.2, clustering efficiency for churn prediction rate is evaluated. Clustering efficiency of SwiftRule is higher than Fast clustering bAsed feature Selection algoriThm (FAST), Secure Mining Protocol and Ranking Adaptation SVM (RA-SVM). Research in Swiftrule has 17.46% higher clustering efficiency than Fast clustering bAsed feature Selection algoriThm (FAST), 8.43% higher clustering efficiency than Secure Mining Protocol and 13.43 % higher clustering efficiency than Ranking Adaptation SVM (RA-SVM).

4.3 Execution Time

Execution time is defined as the time taken to predict the churn based on the expectation count in the telecom industry. It is measured in terms of milliseconds (ms).

$$Execution\ Time\ (ms) = Starting\ Time - Ending\ time\ of\ churn\ prediction$$

Table 4.3 Tabulation of Execution Time for Churn Prediction Techniques

Number of Subscribers (Number)	Execution Time (ms)			
	FAST	Secure Mining Protocol	SwiftRule	RA-SVM
10	28	35	30	39
20	32	39	34	44
30	35	42	37	48
40	38	45	40	53
50	41	49	44	58
60	45	53	48	62
70	48	56	51	65

Execution Time comparison takes place on existing Fast clustering bAsed feature Selection algoriThm (FAST), Secure Mining Protocol, SwiftRule and Ranking Adaptation SVM (RA-SVM).

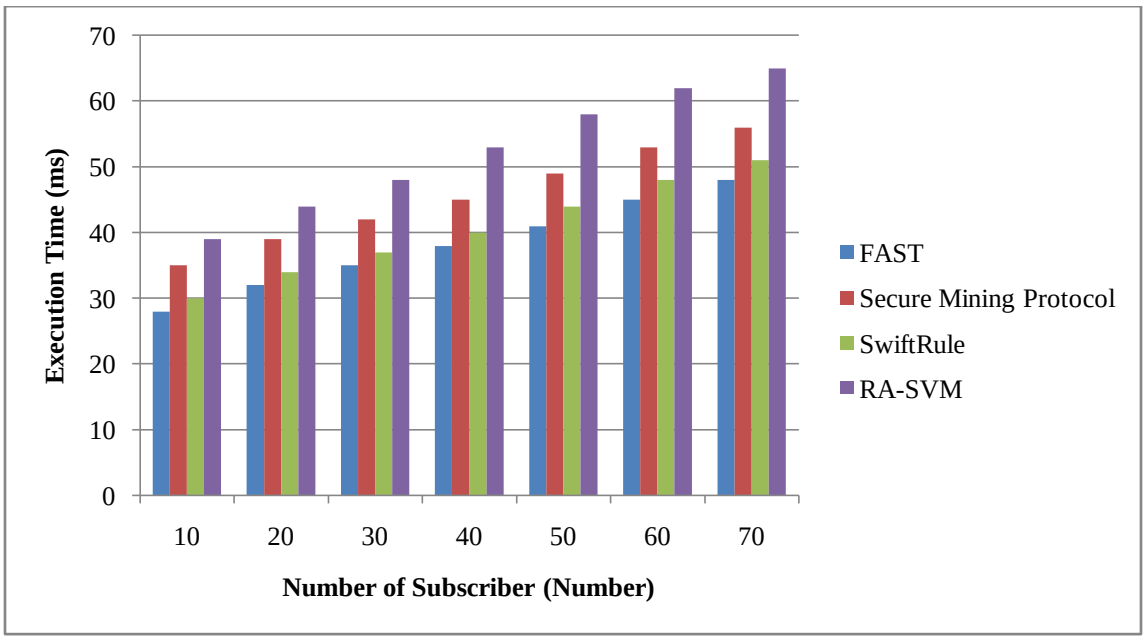

Figure 4.3 Execution Time for Churn Prediction Techniques

From figure 4.3, execution time for churn prediction rate is evaluated. Execution Time of Fast clustering bAsed feature Selection algoriThm (FAST) is comparatively lesser than Swiftrule, Secure Mining Protocol and Ranking Adaptation SVM (RA-SVM). Research in Fast clustering bAsed feature Selection algoriThm (FAST) consumes 6.32% lesser than SwiftRule, 19.89% lesser execution time than Secure Mining Protocol and 38.29 % lesser execution time than Ranking Adaptation SVM (RA-SVM).

5. DISCUSSION ON LIMITATION OF CHURN PREDICTION

In Fast Clustering-Based Feature Subset Selection (FAST), features are separated into clusters using graph-theoretic clustering methods. The representative feature is connected with target classes for each cluster to form a subset of features. Clustering-based strategy of FAST contains high chance of creating a subset of useful and independent features. However, FAST does not explore various types of correlation measures. Some formal properties of feature space clustering are also not ensured.

Segmentation and piecewise polynomial demonstrating are done quick in just a single sit back arrangement utilizing SwiftRule. Static classifier for single segments and then extends to a dynamic classifier for sequences of segments. It also creates the dynamic classifier from a static one maintains the number of different univariate Gaussians low. SwiftRule by hidden Markov models with Gaussian output distributions for identifying the solution fail to address comprehensibility of rules.

Regularization-based algorithm called ranking adaptation SVM (RA-SVM) constructs an efficient domain-specific ranking model. Blackbox ranking model adaptation requires the predictions. Documents similar with domain-specific feature space contain constant rankings. Ranking Support Vector Machines (Ranking SVM) is the efficient learning to rank algorithms. Though, ranking adaptability fails to quantitatively measure specific target domain results.

Event Characterization and Prediction based on Multivariate reconstructed phase space method (MRPS) was introduced by Wenjing Zhang., and Xin Feng., (2014) identifies the multi-dimensional data sequence problems. MRPS discovers the relationship with defined events in the target data sequence but alternative clustering method is not adopted for churn analysis. More complex event problem function cannot be worked out for different applications through this method. Massimiliano Albanese., et al., (2014) explains Top-k unexplained Sequences in Time-Stamped Observation Data that perform the pruning with high search effect. This model fails to predict the system with the high scalable result.

Ying Zhang., et al., (2014) described about Generalized Borda Count (GBC) Approach with Multivalve Object faults which prefers objects to rank and identifies the wide range of fault causing. GBC developed an efficient computational method with comprehensive cost analysis on identifying the faults with pruning techniques. An index based algorithm calculates the rank of the objects. Aggregation methods to compute the final top-k ranking results is not carried out with decision tree.

6. CONCLUSION

In this survey, comparison of different existing churn prediction techniques are carried out in brief manner. From the survey, it can be understood that decision tree based techniques, Neural Network based techniques and regression techniques are used in customer churn. Decision trees are the most common methods used in predicting and evaluating the classification of customer churn problems. The future direction of churn prediction in telecommunication industry is to cluster the churns using churn clustering techniques. In addition to that, perform rule mining operation using fuzzy rule sets and predict the churn using the decision tree with high scalability rate.

REFERENCES

[1] Fosca Giannotti., Laks V. S. Lakshmanan., Anna Monreale., Dino Pedreschi., and Hui (Wendy) Wang., "Privacy-Preserving Mining of Association Rules From Outsourced Transaction Databases", IEEE Systems Journal, Vol. 7, No. 3, September 2013.

[2] Fabrizio Angiulli., Stefano Basta., Stefano Lodi., and Claudio Sartori., "Distributed Strategies for Mining Outliers in Large Data Sets", IEEE Transactions on Knowledge and Data Engineering, Vol. 25, No. 7, July 2013.

[3] Ying Zhang., Wenjie Zhang., Jian Pei., Xuemin Lin., Qianlu Lin., and Aiping Li., "Consensus-Based Ranking of Multivalued Objects: A Generalized Borda Count Approach", IEEE Transactions on Knowledge and Data Engineering, Vol. 26, No. 1, January 2014.

[4] Wenjing Zhang., and Xin Feng., "Event Characterization and Prediction Based on Temporal Patterns in Dynamic Data System", IEEE Transactions on Knowledge and Data Engineering, Vol. 26, No. 1, January 2014.

[5] Qinbao Song., Jingjie Ni., and Guangtao Wang., "A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data", IEEE Transactions on Knowledge and Data Engineering, Vol. 25, No: 1, January 2013.

[6] Tamir Tassa., "Secure Mining of Association Rules in Horizontally Distributed Databases", IEEE Transactions on Knowledge and Data Engineering, Vol. 26, No. 4, April 2014.

[7] Dominik Fisch., Thiemo Gruber., and Bernhard Sick., "SwiftRule: Mining Comprehensible Classification Rules for Time Series Analysis", IEEE Transactions on Knowledge and Data Engineering, Vol. 23, No. 5, May 2011.

[8] Bo Geng., Linjun Yang., Chao Xu, and Xian-Sheng Hua., "Ranking Model Adaptation for Domain-Specific Search", IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No. 4, April 2012

[9] Massimiliano Albanese., Cristian Molinaro., Fabio Persia., Antonio Picariello., and V.S. Subrahmanian., "Discovering the Top-k Unexplained Sequences in Time-Stamped Observation Data", IEEE Transactions on Knowledge and Data Engineering, Vol. 26, No: 3, March 2014

[10] Kai Zheng., Zi Huang., Aoying Zhou., and Xiaofang Zhou., "Discovering the Most Influential Sites over Uncertain Data: A Rank-Based Approach", IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No. 12, December 2012

