Journal of Theoretical and Applied Information Technology
31st May 2015. Vol.75. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

FEATURE SELECTIONS AND CLASSIFICATION MODEL FOR CUSTOMER CHURN

[1]SYAFIQAH NAFIS, [2]MOKHAIRI MAKHTAR, [3]MOHD KHALID AWANG,
[4]M. NORDIN A.RAHMAN, [5]MUSTAFA MAT DERIS

[1, 2, 3, 4] Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, Kuala Terengganu, Malaysia
[5] Faculties of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Malaysia

E-mail: [1] nsyafiqahmnafis@gmail.com, [2] mokhairi@unisza.edu.my, [3] khalid@unisza.edu.my, [4] mohdnabd@unisza.edu.my, [5] mmustafa@uthm.edu.my

ABSTRACT

As customers actively exercise their right to change to a better service and since engaging new customers is more costly compared to retaining loyal customers, customer churn has become the main focus for one organization. This phenomenon affects many industries such as telecommunication companies which need to provide excellent service in order to win over the competition. Several models were developed in previous research using various methods such as the conventional statistical method, decision tree based model and neural network based approach in predicting customer churn. Several experiments were conducted in this research for feature selection and classification from selected customer churn dataset to compare its usefulness among the different feature selections and classifications using a data mining tool. The results from the experiments showed that the Logistic Model Tree (LMT) method is the best method for this dataset with a 95% accuracy enhanced using neural network from previous research.

Keywords: *Customer Churn, Classification, Feature Selection, Telecommunications, Data Mining*

1. INTRODUCTION

Telecommunications is the number one requirement today of the young and the elderly due to advances in computer and network technologies. The telecommunications industry has grown well due to this phenomenon. However, there is stiff competition amongst the telecommunication companies since customers actively exercise their right to switch from one provider to another which satisfies their needs defined as, *churn*. Churn is defined as turnover action by a customer of a business or service. There are also common clashes whereby customers demand excellent service with a cheaper price while are strictly focused on their business goal i.e. to increase the churn rate. The fact is that acquiring a new customer is more expensive as compared to retaining existing subscribers as mentioned in previous research [1].

Econsultancy is the leading source of independent advice and insight on digital marketing and e-commerce founded in 1999. *Econsultancy* published the Cross-channel Marketing Report 2013 which denied some compelling statistics for those who believed that customer retention is undervalued as illustrated in Figure 1 and Figure 2. Figure 1 illustrates a company's concern on relationship marketing. It shows that only 30% of its staff is committed to customer relationship management and about 22% do not really care about relationship marketing which in turn indicates that their awareness of customer relationship is relatively minimal. Figure 2 shows that 70% of the respondents agree that it is cheaper to retain customers than to acquire new customers which supports [1].

Journal of Theoretical and Applied Information Technology
31st May 2015. Vol.75. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

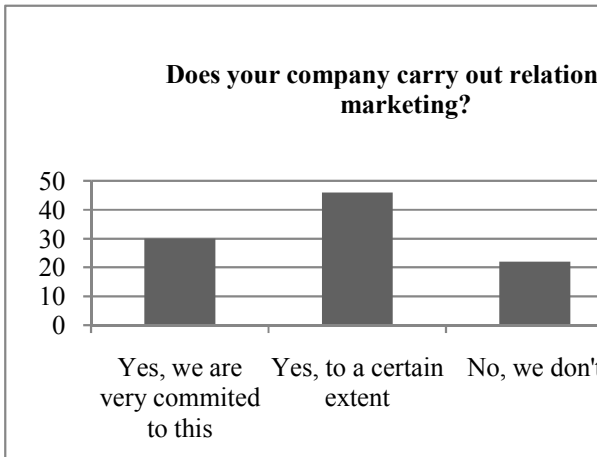

Figure 1.Company Awareness on Relationship Marketing

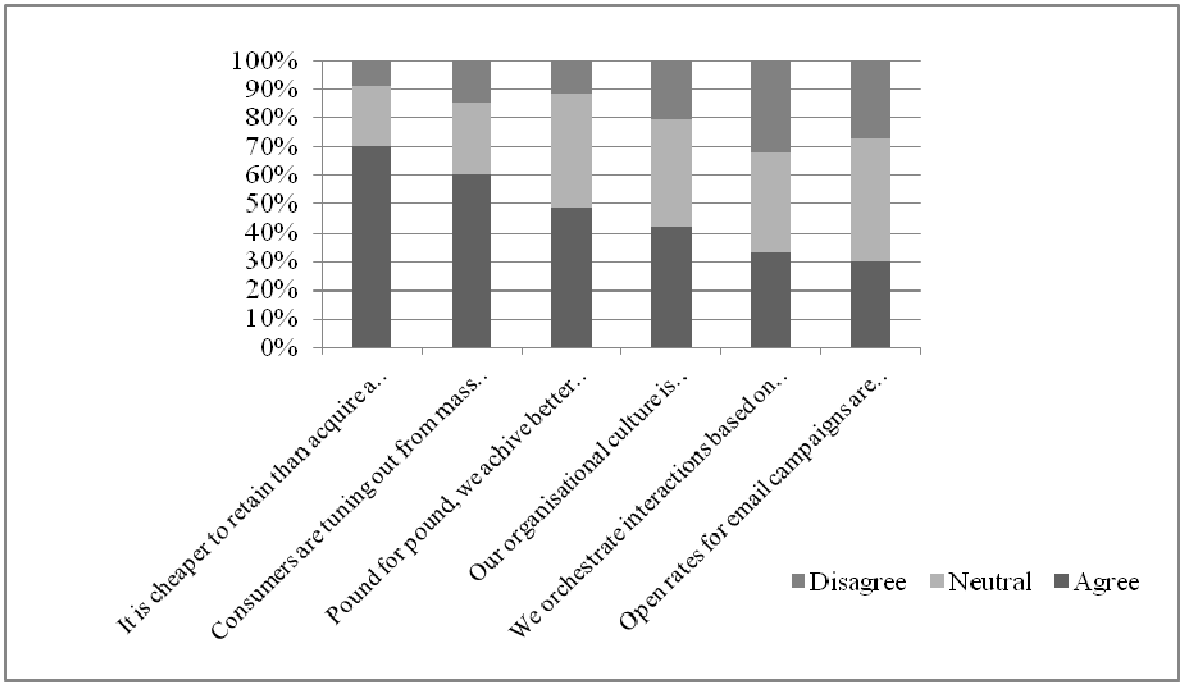

Figure 2.Customer Relationship Management Awareness.

Data Mining (DM) has been proposed as a technique to predict customer churn. Decision tree, neural network and Naïve Bayes are a few examples of classification methods in data mining. Classification refers to a process in data mining for pattern extraction and categorization. Rapid developments in the classification field in recent years have triggered researchers to develop a model to overcome the churn problem in s. Past research has proven that classification is capable of assisting Customer Relationship Management (CRM) concerning churn management as in [2]. CRM is a huge concern in this issue to optimize their marketing strategies or take corrective action to prevent customer churn as well as to maintain and improve their services. In short, if CRMs are fully aware which customers are at high risk of churn, they can design a treatment program to address the issue. For instance, CRMs can offer new telecommunication packages or services to retain

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

existing customers as well as attract new subscribers.

In this paper, it targets to evaluate usefulness among different feature selections and classifications in order to identify the best model to classify customer churn, besides improving accuracy using Neural Network in [3] by demonstrating several experiments for feature selection and classification from selected customer churn dataset.

This paper is organized as follows: In section 2, it describes related work in customer churn, feature selection and classification. Section 3 gives detailed explanation on the tested dataset, software tool used and classification method applied. Section 4 discusses the experiment and analysis followed by the conclusion and future works in section 5.

2. LITERATURE REVIEW

2.1 Related Work

Recently customer churn is the main concern for many companies to maximize profit. Investment in CRM technology has become crucial because of the competitive growth of telecommunication companies which has reached the stage of maturity. Besides, many highly competitive organizations understand that retaining existing and valuable customers is their core managerial strategy to survive in the industry [4]. Churn can be divided into two types which is voluntary churn and involuntary churn. Voluntary churn is varied and more complex. For example, due to the poor service provided, a customer decides to terminate the service. Meanwhile, if a customer is disconnected by the service provider for fraud or non-payment, it is categorized as involuntary churn. The focus in this paper is to overcome voluntary churn because voluntary churn is an unexpected occurrence.

Based on a thorough study in churn management, it was found that a large number of studies have been conducted in various areas. Thus, this research identified which method is the best for customer churn classification. Table 1 below, explains briefly the various methods used in customer churn prediction in different domains.

Table 1.Previous Research in Different Domains

Case Study	Method
Cellular Network Services [3]	Neural Network Based approach
Insurance Customer [5]	Churn Index and Decision Tree CART.
Telecommunications Industry [6]	Decision Tree Analysis
Food Industry [7]	Random Forest and Boosted Trees Techniques
Telecommunication Company [8]	Neural Network (Back Propagation)
Mobile Telephony Industry [9]	Naïve Bayes and Bayesian Network
Wireless Industry [10]	Naïve Bayes
Australia Nursing Industry [11]	Qualitative Design
Bank Customer (India scenario) [12]	CART, C 5.0

2.2 Classification Model

Generally, classification can be referred to as a process to categorize objects according to the characteristics of the objects. Classification is part of the data mining process. In data mining, classification is defined as an analyzing task for a set of pre-classified data objects to study a model or function that can be used or applied to unseen data objects before being placed into one of several predefined classes [13]. Classification has been widely used in many domains such as the application of classification in multimedia data management [14], in email classification [15] and images classification [16]. A thorough search of classification methodologies used in literature revealed that no one methodology can be served as a recipe for a comparison of all computational classification methods [17].

2.3 Feature Selection Method

Feature selection is one of the important phases in data pre-processing for data mining. The objective of feature selection is to select the best subset containing the least number of dimensions that contribute the most to accuracy, and discards the remaining, unimportant dimensions [18]. Basically, two approaches are used in feature selection that is the forward selection and backward selection. This paper uses the forward selection in the Best First feature selection method. Two feature selection evaluators are used. They are the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

CfsSubsetEvaluator and the Leave One Out evaluator. The two feature sets are defined as follows:

a) CfsSubsetEvaluator

Phone, International Plan, Day Min and Customer Service Call

b) Leave One out Evaluator

Voice Mail Plan, Day Min, Eve Min and Customer Service Call

3. CASE STUDY

3.1 WEKA

WEKA was selected as a DM tool for feature selection and classification because of its simple interface as used in [19]. It is developed on Java platform which provides a collection of machine learning and data mining algorithms for data classification, clustering, association rule, and evaluation [20]. WEKA is a software which is distributed under the GNU General Public License, followed by some lessons learned over a period spanning its development and maintenance [21]. In this paper, the WEKA tool was selected for feature selection and classification due to its open source nature and ease of use of its interface which allows the user to apply data mining methods directly to the selected dataset.

3.2 Dataset

The challenge in finding dataset is business confidentiality and privacy. Thus, this paper used a dataset acquired from UCI Repository of Machine Learning Database of the University of California Irvine. The classification is implemented on WEKA, a tool for data mining. The dataset consists of 3,333 cleaned objects and 20 instances along with one indicator whether or not to churn. The instances are briefly explained as follows:

- State: Categorical variable, for the 50 states and the district of Columbia
- Account length: Integer-valued variable for how long account has been active
- Area code: Categorical variable
- Phone number: Essentially a surrogate key for customer identification
- International Plan: Dichotomous categorical having yes or no value
- Voice Mail Plan: Dichotomous categorical variable yes or no value
- Number of voice mail messages: Integer-valued variable
- Total day minutes: Continuous variable for number of minutes customer has used the service during the day
- Total day calls: Integer-valued variable
- Total day charge: Continuous variable based on foregoing two variables
- Total evening minutes: Continuous variable for number of minutes customer has used the service during the evening
- Total evening calls: Integer-valued variable
- Total evening charge: Continuous variable based on previous two variables
- Total night minutes: Continuous variable for storing minutes the customer has used the service during the night
- Total night calls: Integer-valued variable
- Total night charge: Continuous variable based on foregoing two variables
- Total international minutes: Continuous variable for minutes customer has used service to make international calls
- Total international calls: Integer-valued variable
- Total international charge: Continuous variable based on foregoing two variables
- Number of calls to customer service: Integer-valued variable

In this research, 10 fold-cross validations were used. The dataset was tested using four different classification methods which include:

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

decision tree, decision rules, Naïve Bayes and 4 neural networks. These four types of methods were chosen because they are commonly used in data mining.

3.3.1 Decision Tree

Decision tree is one of the well-known classifiers among researchers. It creates a decision tree from data to classify a class. A decision-tree model classifies an instance by sorting it through the tree to the appropriate leaf node, i.e. each node represents a classification [4]. A common decision tree learning algorithm adopts a top-down recursive divide-and conquer strategy to construct a decision tree [13]. Figure 3 shows a simple example of a decision tree.

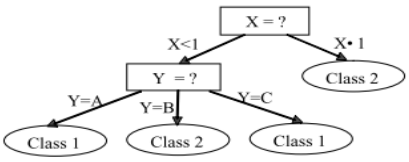

Figure 3. Example Of A Decision Tree

It starts from a root node representing the whole training data at the top. The data is split into two or more subsets (nodes) based on the values of a chosen attribute. This results in a decision tree in which data is partitioned into several nodes or leaves along branches. For each subset a child node is created and the subset is associated with the child at the bottom. Then the process is repeated separately on the data in each of the child nodes, and so on, until a termination criterion is satisfied. Many decision tree learning algorithms exist such as Random Tree, Random Forest and C4.5.

3.3.2 Decision Rules

Decision rules are a set of simple if-then rules. The example of decision rule is "if *age* < *9* and *gender* = *L*, then the example is classified as Class X". Thus, this proves that it is an expressive, yet simple and human readable classification. Examples of decision rules are RIPPER and PART.

3.3.3 Naïve Bayes

The Naïve Bayes classifier is a simple classifier to use and efficient to learn because it is based on Bayes' theorem [13].

Bayes' Theorem,

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$$

Naïve Bayes just requires considering one attribute in each class separately from the training data, thus making it easier and faster. Naïve Bayes always competes well with more sophisticated classifiers and is one of the most effective and efficient inductive learning algorithms for data mining along with machine learning [22].

3.3.4 Neural Network

Neural network also refers to artificial neural network. It simply means a study to mimic the human brain although the human brain is much more complex compared to the artificial neural network developed so far. The neural network technique consists of developing a computer based decision support system to perform as a human brain. Previous research on neural network showed better performance in predicting diseases in the medical field [23]. But Multilayer Perceptron is the simplest and commonly used neural network architecture.

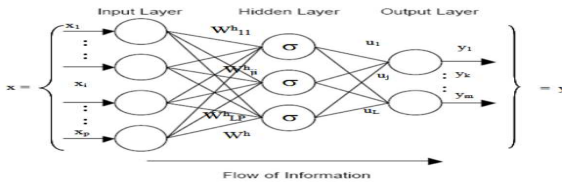

Figure 4. Example of a Neural Network

4. RESULTS AND DISCUSSION

Few experiments have been conducted to find the best feature selection and classification for experimented datasets with 10-fold cross validation. There are four classification techniques used with different feature selections and classifiers, which include Decision Tree, Naïve Bayes, Decision Rules and Neural Network.

Accuracy is used to evaluate the results from the classifiers. Table 2 lists the results for decision tree classifications. Results show that the J48 method gives the highest accuracy of 94.03 % without any feature selection. Accuracy for all decision tree methods decreases or remains the same if feature selection is included in the experiment. This result indicates that feature selection is not needed in

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

decision tree classification for the selected dataset because decision tree classification comprises a flowchart like structure whereby removing one attribute could affect the flowchart.

Table 3 above lists Naïve Bayes classification methods. Bayes Net with feature selection (a) gives the highest accuracy of 87.97 %. But the rest of the methods decrease in accuracy with feature selection. At this point, it can be concluded that Naïve Bayes classification methods do not need feature selection beforehand because Naïve Bayes classifiers are based on the probability model.

Table 4 shows the classification methods for Decision Rule. JRip without feature selection provides the highest accuracy (94.99%). As in Decision Tree classification, add-in feature selection in Decision Rule classification reduces accuracy. This shows that all instances are important for classifying with regards to Decision Rule classifiers

Table 2. Decision Tree Classification

Name of Method	Decision Stump	Hoeffding Tree	J48	Random Forest	Random Tree	REPTree	LMT
No feature selection accuracy (%)	86.56	87.76	94.03	87.01	86.17	85.51	93.88
Accuracy with feature selection (a) (%)	86.56	86.35	88.63	85.51	85.84	85.51	88.81
Accuracy with feature selection (b) (%)	86.56	86.65	91.81	91.36	86.86	92.05	92.14

Table 3. Naïve Bayes Classification

Name of Method	Bayes Net	Naïve Bayes	Naïve Bayes Updateable
Accuracy without feature selection (%)	85.99	87.64	87.64
Accuracy with feature selection (a) (%)	87.97	86.56	86.56
Accuracy with feature selection (a) (%)	87.25	86.80	86.80

Table 4. Decision Table Classification

Name of Method	Accuracy without feature selection (%)	Accuracy with feature selection (a) (%)	Accuracy with feature selection (b) (%)
Decision Table	90.15	88.63	90.25
JRip	94.99	88.51	91.63
OneR	85.51	85.59	86.38
PART	90.49	88.60	91.51

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 5. Various classification methods with state and phone attributes removed

Type of classification	Name of Method	Accuracy (%)
Neural Network	Multilayer Perceptron	93.85
Neural Network	Voted Perceptron	85.57
Decision Tree	J48	94.42
Decision Tree	LMT	95.32
Decision Rule	PART	94.09
Decision Rule	JRip	95.08
Naïve Bayes	Naïve Bayes	88.30
Naïve Bayes	Naïve Bayes Updateable	88.30

Table 5 summarizes the best method from various classification techniques. This experiment is based on [3] which removed *state* and *phone* since these variables were used for identification purposes only. Thus, this step can be considered as the feature selection. No feature selection method in WEKA is used. Logistic Model Trees (LMT) achieved the highest accuracy of 95.32%. In short, it can be concluded that feature selection is needed but selecting the right feature selection is crucial. Therefore, this calls for future research.

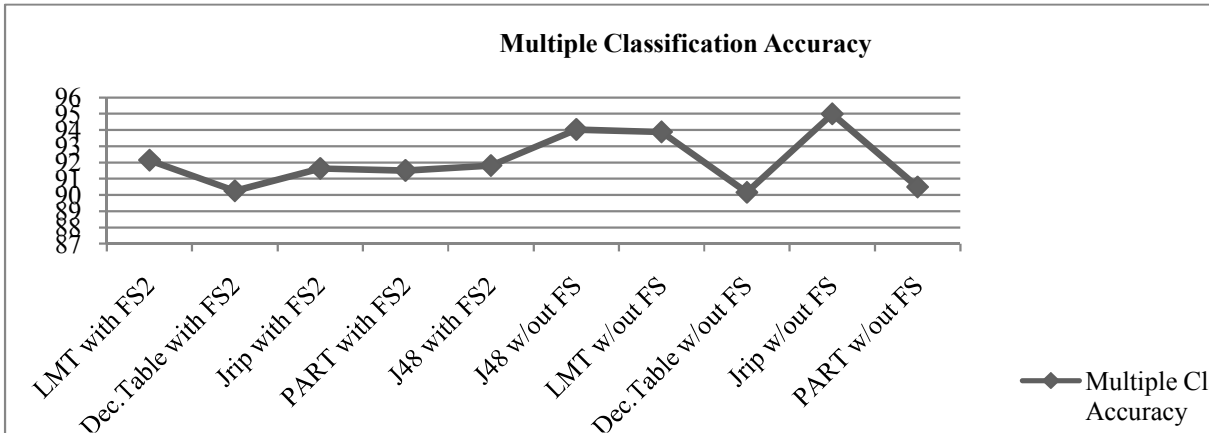

Figure 5. Multiple Classification Accuracy using Different Feature Selections

Figure 5 and Figure 6 summarize the best classification methods among the four classifiers. Figure 1 shows accuracy in having a feature selection (b) and not having a feature selection (b). Figure 2 shows the best classification method when attributes such as state and phone are removed. LMT classifier secured the highest accuracy among the other five classifiers.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

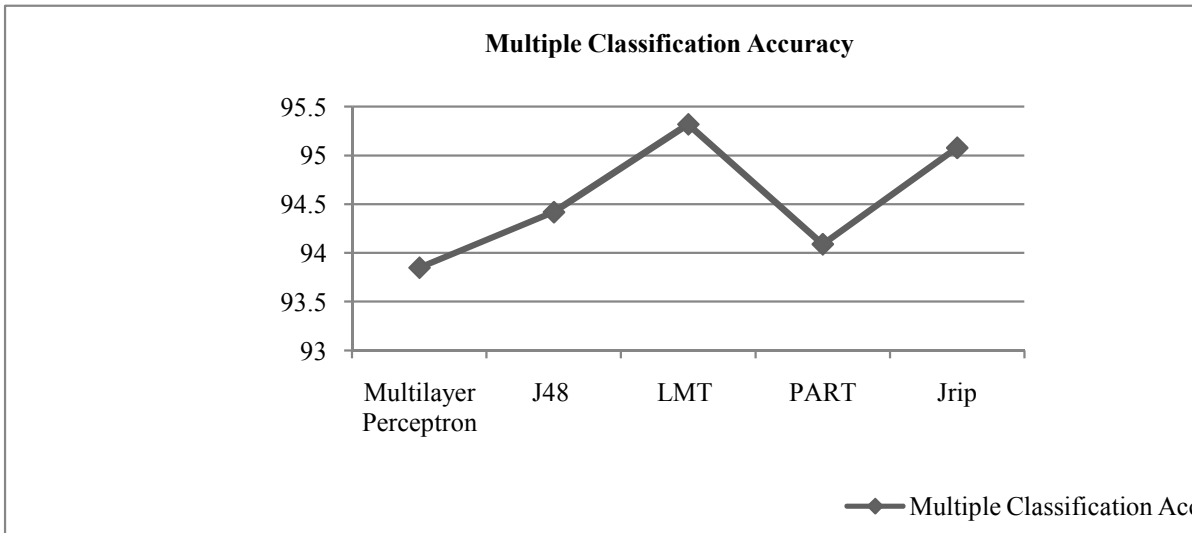

Figure 6. Multiple Classification Accuracy with 'state' and 'phone' attributes removed

Table 6. Confusion Matrix for JRip Classifier

JRip Classifier			
True Label	False	True	Totals
False	2806	44	2850
True	120	363	483
Total	2926	407	3333

Table 7. Confusion Matrix for LMT Classifier

LMT Classifier			
True Label	False	True	Totals
False	2811	39	2850
True	117	366	483
Total	2928	405	3333

Table 6 and Table 7 show two different confusion matrixes for LMT and JRip. The LMT classifier classified about 90% churners while JRip was able to predict 89% churners. In short, the LMT is relatively better in predicting customer churn. Hence, it is the best classification model for this dataset but at the same time the LMT is also good for classifying non-churners with 96% accuracy. However, it is crucial to predict the churners as compared to the non-churners class as the objective of the company is to maximize profit by focusing on churners. For the non-churners, error in classifying them would not affect the CRM or the overall company profit.

5. CONCLUSION

Churn prediction is crucial not only in the telecommunications industry since customers are aware of their right to choose the best service.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Thus, having the best prediction model would assist in retaining loyal customers in CRM.

For the tested dataset, it can be concluded that it cannot represent the current telecommunications phenomenon and do not fit in with the case in Malaysia even though the Logistic Model Tree gives the best accuracy of 95% for tested dataset. The result would be different if different datasets were tested using the same method. For example, the latest and current mobile telecommunication phenomenon would have different features, which would impact on the feature selection sets for example, its mobile data usage and roaming usage.

For future work, several issues could be considered. First, careful attention needs to be exercised when choosing feature selection in order to improve churn accuracy and that only important features are selected. For example, its application in soft computing such as Rough Set Theory needs to be considered. Secondly, it is a huge success if a model can predict churners beforehand although it is better to know when customers churn. Throughout this research, data limitation is identified. Experimented dataset do not represent current situation in Malaysia since it is retrieved from established repository due to business confidentiality and privacy. Finally, models can be tested for its fit for purpose with the latest/current telecommunications or mobile network sector

ACKNOWLEDGEMENTS

This work is partially supported by Universiti Sultan Zainal Abidin (UniSZA) and The Ministry of Education, Malaysia (Grant No.FRGS/2/2013/ICT07/UniSZA/02/2).

REFERENCES:

[1] J. Lu and D. Ph, "Predicting Customer Churn in the Telecommunications Industry — An Application of Survival Analysis Modeling Using SAS â," 2011.

[2] J. Hadden, A. Tiwari, R. Roy, and D. Ruta, "Churn Prediction using Complaints Data," no. 1999, 2006.

[3] A. Sharma, "A Neural Network based Approach for Predicting Customer Churn in Cellular Network Services," vol. 27, no. 11, pp. 26–31, 2011.

[4] C.-F. Tsai and M.-Y. Chen, "Variable selection by association rules for customer churn prediction of multimedia on demand," *Expert Syst. Appl.*, vol. 37, no. 3, pp. 2006–2015, Mar. 2010.

[5] R. A. Soeini and K. V. Rodpysh, "Applying Data Mining to Insurance Customer Churn Management," vol. 30, pp. 82–92, 2012.

[6] N. A. Haris, O. Support, and S. Division, "Data Mining in Churn Analysis Model for Telecommunication Industry," vol. 1, no. 19, pp. 19–27, 2010.

[7] S. Nabavi and S. Jafari, "Providing a Customer Churn Prediction Model Using Random Forest and Boosted Trees Techniques (Case Study: Solico Food Industries Group)," vol. 3, no. 6, pp. 1018–1026, 2013.

[8] R. J. Jadhav, "Churn Prediction in Telecommunication Using Data Mining Technology," vol. 2, no. 2, pp. 17–19, 2011.

[9] C. Kirui, L. Hong, W. Cheruiyot, and H. Kirui, "Predicting Customer Churn in Mobile Telephony Industry Using Probabilistic Classifiers in Data Mining," vol. 10, no. 2, pp. 165–172, 2013.

[10] S. V Nath, "Customer Churn Analysis in the Wireless Industry: A Data Mining Approach Customer Churn Analysis in the Wireless Industry: A Data Mining Approach," no. 561, pp. 1–20, 2003.

[11] A. J. Dawson, H. Stasa, M. a Roche, C. S. E. Homer, and C. Duffield, "Nursing churn and turnover in Australian hospitals: Nurses perceptions and suggestions for supportive strategies.," *BMC Nurs.*, vol. 13, p. 11, 2014.

[12] U. D. Prasad, "Prediction of Churn Behavior of Bank Customers," *Bus. Intell. J.*, vol. 5, pp. 96–101, 2012.

[13] A. An, "Classification Methods," pp. 144–149, Jun. 2005.

[14] M. N. A. Rahman, Y. M. Lazim, and F. Mohamed, "Applying Rough Set Theory in Multimedia Data Classification," vol. 1, no. 3, pp. 683–693, 2011.

[15] Z. Z. Z. Zhu, "An Email Classification Model Based on Rough Set and Support Vector Machine," *2008 Fifth Int. Conf. Fuzzy Syst. Knowl. Discov.*, vol. 5, pp. 403–408, 2008.

[16] N. S. Kamarudin, M. Makhtar, S. A. Fadzli, M. Mohamad, F. S. Mohamad, M. F. Abdul Kadir, "Comparison of Image Classification Techniques Using Caltech 101 Dataset," vol. 71, no. 1, 2015.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[17] Q. A. Al-radaideh, "The Impact of Classification Evaluation Methods on Rough Set Based Classifier," no. 1, pp. 2–6, 2008.

[18] L. Ladha and T. Deepa, "Feature Selection Methods and Algorithms," *Int. J. Comput. Sci. Eng.*, vol. 3, pp. 1787–1797, 2011.

[19] P. Ozer, "Data Mining Algorithms for Classification," no. January, 2008.

[20] A. A. A. Hafieza Ismail, Fadhilah Ahmad, "Seminar Penyelidikan Siswazah UniSZA Peringkat Kebangsaan (SEMPSIS), Implementing WEKA as a Data Mining Tool to Analyze Students ' Academic Performances using Naïve Bayes Classifier Nur Hafieza Ismail , Fadhilah Ahmad , Azwa Abdul Aziz University Sultan," no. July 2011, 2013.

[21] M. A. Hall and I. H. Witten, "WEKA — Experiences with a Java Open-Source Project," *J. Mach. Learn. Res.*, vol. 11, pp. 2533–2541, 2010.

[22] A. M. Taha, A. Mustapha, and S. Der Chen, "Naive Bayes-guided bat algorithm for feature selection," *Sci. World J.*, vol. 2013, 2013.

[23] D. R. Chowdhury, M. Chatterjee, and R. K. Samanta, "An Artificial Neural Network Model for Neonatal Disease Diagnosis," *Int. J. Artif. Intell. Expert Syst.*, vol. 2, pp. 96–106, 2011.