Improving Accuracy and Performance of Customer Churn Prediction Using Feature Reduction Algorithms

Mohd Khalid Awang, Mokhairi Makhtar, Mohd Nordin Abdul Rahman
Faculty of Informatics and Computing,
University Sultan Zainal Abidin (UniSZA)
Tembila, Besut, Terengganu, Malaysia.
khalid@unisza.edu.my

***Abstract*—Prediction of customer churn is one of the most essential activities in Customer Relationship Management (CRM). However, the state-of-the-art of the customer churn prediction approach only focuses on the classifier selection in improving the accuracy and performance of churn prediction, but rarely contemplate the feature reduction algorithms. Furthermore, there are numerous attributes that contribute to customer churn and it is crucial to determine the most substantial features in order to acquire the highest prediction accuracy and to improve the prediction performance. Feature reduction decreases the dimensionality of the information and may allow learning algorithms to function faster and more effectively and able to produce predictive models that deliver the highest rate of accuracy. In this research, we investigated and proposed two (2) different feature reduction algorithms which are Correlation based Feature Selection (CFS) and Information Gain (IG) and built classification models based on three 3) different classifiers, namely Bayes Net, Simple Logistic and Decision Table. Experimental results demonstrate that the performance of classifiers improves with the application of features reduction of the customer churn data set. A CFS feature reduction algorithm with the Decision Table classifier yields the highest accuracy of 92.08% and has the lowest RMSE of 0.2554. This study recommends the use of feature reduction algorithms in the context of CRM for churn prediction to improve accuracy and performance of customer churn prediction.**

***Index Terms*—Feature Reduction; Feature Selection; Customer Churn Prediction.**

I. INTRODUCTION

Prediction of future customers is one of the most significant activities that form the heart for all Customer Relationship Management (CRM) program. [1] Indicated that customer retention leads to improved sales and reduced marketing costs compared to selling to fresh customers. Maintaining the right customer is the determinant of profitability for longer term, rather than finding new clients that will increase the cost [2]. In the telecommunication industry, the issue of customer retention and loyalty management is becoming increasingly significant as competition from competitors that increasingly aggressive [3, 4].

A. Customers Churn

The behavior of customers who shift from one service provider to another is known as churn and now has become a major concern of network providers. Finding the prospective churners can help the companies to retain their customers. Many researchers have tried to establish various models to predict customer churn, for example the decision tree, support vector machine (SVM), neural network, genetic algorithm, generalized additive models (GAM), logistic regression and linear regression analysis [5,6,7,8,9,10,11]. Table 1 shows some of the current researches and approaches in customer churn prediction. However, current methods for churn prediction are not effective and need to be improved because there are many factors that contribute to customer churn. The prediction accuracy could be improved by removing some of the irrelevant features through reduction algorithms.

Table 1
Research in Customer Churn Prediction

Researchers	Churn Factors/Features	Classifiers
(Shaaban, Helmy, Khedr, & Nasr, 2012)	Information data Usage data Complaints data	Neural Network SVM Decision Tree
(Wang, Sanguansintukul, & Lursinsap, 2008)	Usage data Billing data Debt data	Neural Network
(Huang et al., 2009)	Demographic data Billing data Usage data Information grants	Linear Regression Decision Tree Neural Network SVM
(Khan et al., 2010)	Demographic data Billing data Usage data	Decision Tree Logistic Regression Neural Network
Coussement et al., (2015)	Demographic data Billing data Usage data	Generalized Additive Models (GAM). Logistic regression

B. Feature Selection Algorithms

Feature selection has been applied in the various domains such as marketing, business application, pattern recognition, image processing, classification and prediction. When dealing with a real application, it is normal to encounter a large number or data set and enormous size of features. In most of the cases, only some of the features are significant and relevant to the task. The rest of the features are considered irrelevant and insignificant; thus, it will not only reduce the performance, but at the same time reduce the classification accuracy. Therefore, selecting an appropriate and small feature subsets of the original features not only helps to overcome the "curse of dimensionality" but also important to increase the performance and accuracy of the classification [12].

The aim of feature selection is to find a feature subset that has the most discriminating information from the original feature set. Since there are numerous attributes that contribute

to customer churn and it is crucial to determine the most substantial features in order to acquire the highest prediction accuracy, this paper aims to investigate and proposed two (2) different feature reduction algorithms which are Correlation based Feature Selection (CFS) and Information Gain (IG) and built classification models based on three (3) different classifiers, namely Bayesian Network, Simple Logistic and Decision Table.

a. Correlation based Feature Selection (CFS)

CFS evaluates the substance of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. The aim of CFS is to discover a subset of features that are highly connected with the class while having low inter correlation. CFS assumes that features are conditionally independent and it able to ascertain relevant features when moderate feature dependencies exist [13].

b. Information gain (IG)

The information gain ratio is the ratio between the information gain and the intrinsic value. The information gain measure is used to select the test attribute of each node of the decision tree classification. IG evaluates the significance of an attribute by measuring the information gain with respect to the class. The information gain ratio is a modification of the information gain that reduces its bias by taking the number and size of branches into account when choosing the significant attributes [14]. The attribute with the highest gain ratio is selected as the splitting attribute implies that it is the most significant attribute.

C. Classification Models

Classification is a basic task in data analysis and data mining that requires the construction of a classifier. A classifier is a function that assigns a class label to the instances described by a set of attributes. The induction of classifiers from data sets of pre classified instances is a central problem in machine learning. Numerous approaches to this problem are based on various functional representations such as decision trees, decision lists, neural networks, decision graphs, and rules. In this study, we explore three (3) different classifiers, namely Bayesian Network, Simple Logistic and Decision Table.

a. Bayesian Network

A Bayesian Network classifier is based on graph structure that lets us to represent and reason about an ambiguous domain [15]. Weka implements Bayesian Network classifier using various search algorithms and quality measures. It also provides data structures such as the network structure and the conditional probability distributions.

b. Simple Logistic

Simple logistic algorithm in Weka is a classifier for building linear logistic regression models introduced by Summer et all. [16]. It uses LogitBoost with simple regression functions as base learners for fitting the logistic models. The optimal number of LogitBoost iterations to perform is cross-validated, which leads to automatic attribute selection.

c. Decision Table

Decision table algorithm in Weka is based on a simple decision table majority classifier developed by Kohavi [17]. A decision table is a machine learning algorithm that relies on a hierarchical table to break down the data with each level having two attributes. Decision table is a classification model used to make predictions and behave like decision trees or neural nets.

II. EXPERIMENTAL

The data set for this research is based on the customer churn taken from the UCI Repository of Machine Learning Databases. The dataset consists of 3,333 cleaned objects and 20 instances along with one indicator whether or not to churn. The original dataset is represented in table 2. In this research, we investigated and proposed two (2) different feature reduction algorithms which are Correlation based Feature Selection (CFS) and Information Gain (IG) and built classification models based on three 3) different classifiers, namely Bayes Net, Simple Logistic and Decision Table.

Table 2
Original Features of Customer Churn Prediction

Input Features	Data Type	Description
X_1=State	Categorical	Represent the 50 states and the district of Columbia
X_2=Account length	Numeric	The variable for how long account has been active
X_3=Area code	Categorical	Represent the area code
X_4=Phone number	Text	Essentially a surrogate key for customer identification
X_5=International Plan	Categorical	Dichotomous categorical having yes or no value
X_6=Voice Mail Plan	Categorical	Dichotomous categorical variable yes or no value
X_7=Number of voice mail messages	Numeric	Integer valued variable
X_8=Total day minutes	Numeric	Continuous variable for number of minute customer has used the service during the day
X_9=Total day calls	Numeric	Integer-valued variable
X_{10}=Total day charge	Numeric	Continuous variable based on foregoing two variables
X_{11}=Total evening minutes	Numeric	Continuous variable for number of minute customer has used the service during the evening
X_{12}=Total evening calls	Numeric	Integer-valued variable
X_{13}=Total evening charge	Numeric	Continuous variable based on previous two variables
X_{14}=Total night minutes	Numeric	Continuous variable for storing minutes the customer has used the service during the night
X_{15}=Total night calls	Numeric	Integer-valued variable
X_{16}=Total night charge	Numeric	Continuous variable based on foregoing two variables
X_{17}=Total international minutes	Numeric	Continuous variable to minute customer has used service to make international calls
X_{18}=Total international calls	Numeric	Integer-valued variable
X_{19}=Total international charge	Numeric	Continuous variable based on foregoing two variables
X_{20}=Number of calls to customer service	Numeric	Integer-valued variable
Y_1=actual result	Categorical	OUTPUT-Churn: Yes or No

III. RESULTS AND DISCUSSION

A. Analysis of the Feature Selection Algorithms

Table 3 shows the experimental results by selecting the significant features based on CFS algorithm and Info Gain algorithm. CFS

algorithm considers only 7 features which representing the reduction of 65% of the original features. On the other hand, the Info Gain algorithm selects 11 parameters which shown the reduction of 55% of the original features. Most of the selected features in CFS are also considered important in the Info Gain algorithm.

Table 3
Reduced Features for Customer Churn Prediction

CFS algorithm	Info Gain algorithm
	X_{20}=Number of calls to customer service
X_5=International Plan	X_5=International Plan
X_6=Voice Mail Plan	X_8=Total day minutes
X_8=Total day minutes	X_{10}=Total day charge
X_{11}=Total evening minutes	X_{17}=Total international minutes
X_{17}=Total international minutes	X_{19}=Total international charge
X_{18}=Total international calls	X_6=Voice Mail Plan
X_{20}=Number of calls to customer service	X_7=Number of voice mail messages
	X_{18}=Total international calls
	X_{12}=Total evening calls
	X_{13}=Total evening charge

A. Analysis of the Performance of the Prediction

The main goal of this research is to evaluate the influence of feature selection algorithms on the accuracy and performance of the customer churn prediction. The experiments were performed with the assistance of the Weka machine learning package. We first compare Correlation based Feature Selection (CFS) and Information Gain (IG). The adopted classifiers are Bayes Network, Simple Logistic and Decision Table.

The performance of the classifiers is measured by the prediction accuracy, Root Mean Square Error (RMSE) and the time taken to complete the task. Note that the TP (True Positive) is the number of subscribers predicted to churn who actually churned; FP (False Positive) is the number of subscribers predicted to churn, but did not; FN (False Negative) is the number of subscribers predicted not to churn but did not; TN (True Negative) is the number of subscribers predicted not to churn who actually did. Then the hit ratio TP/(TP+FP), recall ratio TP/(TP+FN).

The Root Mean Square Error (RMSE) (also called the root mean square deviation, RMSD) is a frequently used measure of the difference between values predicted by a model and the values actually observed from the environment that is being modelled. These individual differences are also called residuals, and the RMSE serves to aggregate them into a single measure of predictive power.

Table 4
The Performance of the Prediction

Classifiers	Feature Reduction Algorithms	Prediction Accuracy	Root mean squared error (RMSE)	Time taken to build model (seconds)
Bayes Net	All features	86.7687	0.3082	0.15
	Info Gain Eval	86.7687	0.3082	0.05
	CFS Eval	87.2487	0.2949	0.01
Simple Logistic	All features	85.9886	0.3179	1.06
	Info Gain Eval	86.0186	0.3172	0.16
	CFS Eval	86.0186	0.3172	0.63
Decision Table	All features	90.1590	0.2837	0.30
	Info Gain Eval	90.1590	0.2837	0.20
	CFS Eval	92.0792	0.2554	0.07

Table 4 displays the prediction accuracy of Bayes Net, Simple Logistic and Decision Table classifiers with and without feature selection using Information Gain and CFS on the customer churn dataset. A CFS feature reduction algorithm with the Decision Table classifier produces the highest accuracy of 92.08% and lowest RMSE of 0.2554. Furthermore, it is also efficient in term of time taken to build the model which is only takes 0.07 seconds.

Based on the results, we could claim that feature reduction algorithm is able to improve the prediction accuracy and at the same time producing lower error rates. Furthermore, it also improves the overall performance of the classifiers by reducing the time taken to complete the prediction. In the long run, it is beneficial to the telecommunication companies either for effectiveness in prediction of customer churn or reducing the cost by selecting the smaller size of churn's factors.

IV. CONCLUSIONS

Customer churn prediction is one of the core issues in the telecommunications industry. However, the state-of-the-art of the customer churn prediction approach only focuses on the classifiers in improving the accuracy and performance of churn prediction but rarely consider the feature selection algorithms. Classifiers, which are employed machine learning algorithms are known to degrade in performance and prediction accuracy when faced with many features that are not necessary for rule discovery. In this research, we investigated and proposed two (2) different feature reduction algorithms which are Correlation based Feature Selection (CFS) and Information Gain (IG) and built classification models based on three 3) different classifiers, namely Bayes Net, Simple Logistic and Decision Table. The best prediction model with the prediction accuracy of 92.08% produced by the Correlation based Feature Selection (CFS) method with the Decision Table classifiers. Experiment results show that the performance of classifiers improves with the application of features reduction of the customer churn data set. Overall, this study recommends the use of feature reduction algorithms in the context of CRM for churn prediction to improve accuracy and performance of customer churn prediction. Feature selection techniques show that more information is not always good in machine learning applications.

It should be noted that although there are many other data reduction algorithms available in the literature, this paper only focuses on two (2) different feature reduction algorithms which are Correlation based Feature Selection (CFS) and Information Gain (IG). It is suggested that for future research to conduct a study on other reduction algorithms. In addition, the future work could also consider other learning algorithms such as Neural Network, Support Vector Machine or an ensemble method.

ACKNOWLEDGEMENT

This work is partially supported by UniSZA and KPM (Grant No. FRGS/2/2013/ICT07/UniSZA/02/2).

REFERENCES

[1] Christopher, M., & Peck, H. 2012. *Marketing logistics*. Routledge.
[2] Verhoef, P. C., & Lemon, K. N. 2013. Successful customer value management: Key lessons and emerging trends. *European Management Journal*, 31(1), 1-15.
[3] Ismail, M. R., Awang, M. K., Rahman, M. N. A., & Makhtar, M. 2015. A Multi-Layer Perceptron Approach for Customer Churn Prediction. *International Journal of Multimedia and Ubiquitous Engineering*, 10(7), 213-222.
[4] Awang, M. K., Rahman, M. N. A., & Ismail, M. R. 2012. Data Mining for Churn Prediction: Multiple Regression Approach. *In Computer Applications for Database, Education, and Ubiquitous Computing*. Springer Berlin Heidelberg: 318-324.
[5] Coussement, K., Benoit, D. F., & Van den Poel, D. 2010. Improved marketing decision making in a customer churn prediction context

using generalized additive models. *Expert Systems with Applications*, 37(3): 2132-2143.

[6] De Bock, K. W., & Van den Poel, D. 2011. An empirical evaluation of rotation-based ensemble classifiers for customer churn prediction. *Expert Systems with Applications*, 38(10): 12293-12301.

[7] Shaaban, E., Helmy, Y., Khedr, A., & Nasr, M. 2012. A proposed churn prediction model. *IJERA*, 2: 693-697.

[8] Wang, Y., Sanguansintukul, S., & Lursinsap, C. 2008. The customer lifetime value prediction in mobile telecommunications. *In Management of Innovation and Technology, 2008. ICMIT 2008. 4th IEEE International Conference* on IEEE. 565-569.

[9] Huang, B., Kechadi, M. T., & Buckley, B. 2012. Customer churn prediction in telecommunications. *Expert Systems with Applications*, 39(1): 1414-1425.

[10] Khan, A. A., Jamwal, S., & Sepehri, M. M. 2010. Applying data mining to customer churn prediction in an internet service provider. *International Journal of Computer Applications*, 9(7): 8-14.

[11] Coussement, K., Benoit, D. F., & Van den Poel, D. 2015. Preventing customers from running away! Exploring generalized additive models for customer churn prediction. *In The Sustainable Global Marketplace*. Springer International Publishing; 238-238.

[12] Chandrashekar, G., & Sahin, F. 2014. *A survey on feature selection methods*. Computers & Electrical Engineering, 40(1): 16-28.

[13] Hall, M. A. 1999. *Correlation-based feature selection for machine learning* (Doctoral dissertation, The University of Waikato).

[14] Guyon, I., & Elisseeff, A. 2003. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3: 1157-1182.

[15] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1): 10-18.

[16] Sumner, M., Frank, E., & Hall, M. (2005). Speeding up logistic model tree induction. In *Knowledge Discovery in Databases: PKDD 2005*. Springer Berlin Heidelberg: 675-683.

[17] Kohavi, R. 1995. The power of decision tables. *In Machine Learning*: ECML-95. Springer Berlin Heidelberg: 174-18