Int. J. Advance Soft Compu. Appl, Vol.9, No. 3, Nov 2017
ISSN 2074-8523

Prediction on Customer Churn in the Telecommunications Sector Using Discretization and Naïve Bayes Classifier

Tan Yi Fei[1], Lam Hai Shuan[1], Lai Jie Yan[1]
Guo Xiaoning[1], Soo Wooi King[2]

[1]Engineering Big Data Lab,
Faculty of Engineering, Multimedia University, Malaysia
[2]Faculty of Computing and Informatics, Multimedia University, Malaysia
e-mail: yftan@mmu.edu.my, hslam@mmu.edu.my,
jieyan_lai@hotmail.com, guo.xiaoning@mmu.edu.my,
wksoo@mmu.edu.my

Abstract

In the telecommunications industry, the competitive intensity for retaining existing customers and avoiding losing valuable customers to competitors has increased dramatically. It is a problem of great concern to companies. Customer retention may be boosted by deploying a prediction model to monitor customer activities. In this paper, two experiments with the implementation of data processing techniques using K-Means and Equal-Width Discretization(EWD) combined with Naïve Bayes are performed respectively to conduct a comparison of techniques to identify probable churn activities. Usually, the data generated are of massive size and with high-dimensionality. In order to accommodate fast processing, casual heuristics is a preferred deployment. The technique which integrated different algorithm is implemented using Python language under a single processor environment. By using the correlation between attributes, the experimental results show that this can improve the model in identifying the key factors in churn prediction. The results have demonstrated promising overall accuracy.

Keywords: *Naïve Bayes , K-Means, data discretization, telecommunications churn, prediction*

1 Introduction

The average cost for a new customer acquisition in the telecommunications industry is commonly known to be 10 times higher in comparison to retaining an existing customer[1]. This is due to the highly-saturated and competitive market in the telecommunications industry. According to a survey, a telecommunications carrier will lose approximately 27% of its subscribers each year [2]; hence customer retention is integral to corporate success. The telecommunications industry generates high velocity and high volume data each day. This leads to the pioneering adoption of data mining technology in telecommunications industry in this volatile economy. Over 300 millions of call detail records are produced daily and these call details are kept within the online server for several months. This implied that billions of call details are promptly accessible for data mining purposes [3].

However, as the industry becomes more saturated, there will be several challenges associated in utilizing the telecommunications data. It is crucial to educate operators on the needs for data types, data sources, and requirements for different use cases. A challenging issue faced by the operators is the high customer churn rate. This would highly affect the revenue stream of the telecommunications operators. According to research, it generally takes at least 3.5 years for telecommunications operators to break even on the subscriber acquisition cost (SAC) [4]. Though in reality, the average customer subscription duration is limited to only two years. Therefore, to reduce the churn propensity and the company's overall SAC, it would be advisable for telecommunications companies to find opportunities in the big data analytics space to strategize and monetize for consistent revenue increment. The telecommunications companies have massive amount of customer data obtained from third parties, such as phone bill information and in-depth personal demographic data, for instance: credit score information [2]. The proposed technique integrates the multi-classifier approach to deal with the challenge of a highly-skewed distribution of churn and non-churn customers. The data is collected at individual customer level instead of contract basis level as it is very common for customers to hold multiple service contracts under the same carrier.

Naïve Bayes classifier is from the simple probabilistic classifier family[5]. It is usually used for the dichotomous dataset. Naïve Bayes classifier has been widely used in numerous industries, for instance, the Intrusion Detection System (IDS) for defensive network infrastructure [6], customer churn predictions, segmentation and fraud detection [7]. Other studies were also carried out for dengue outbreak using Naïve Bayes as one of their multiple rule base classifiers [8]. Authors in [13] investigated the dengue's hotspots using Malaysian open data from 2010 until 2015. The findings show that Selangor has the highest dengue outbreak compare to other states due to dense population with highest low-cost residential with improper sanitary and waste management.

In this paper, we implement Naïve Bayes classifier algorithm to identify existing input data and the classifications process for predicting customer churn rate. This model calculates the probability of customers transitioning to another service provider using the customer details. Besides, the extraction of features employed two different unsupervised learning approach; Equal-Width Discretization and K-Means clustering. With the adoption of data discretization during prediction process, the model yields better prediction by achieving lower classification error as claimed in [9].

2 Related Works

G. Vennila et. al. [6] has documented a classification technique that implemented Naïve Bayes using a honeypot to categorise the Voice over Internet Protocol Network(VoIP) deals with DoS (Denial of Service), signature collection, SPIT (Spam over Internet Telephony) to identify patterns of malicious attacks. This approach is designed to segregate the abnormalities within VoIP architecture. The classifier was trained by the data captured at the honeypot for one week. The classifier is then used for predicting the categories of the incoming packets.

A churn prediction had been carried out by IBM Pakistan and IBM United Kingdom [7]. Their works were divided into two phases where the first phase focused on the identification of churn customers using Key Performance Indicators; while in the second phase, the team deployed their model through traditional learning algorithms. The training dataset presented in this paper consists of three main data sources. Most of the data used were continuous data mainly from 6 months of prepaid and 12 months of postpaid transactional data. The team developed three different models to accommodate different packages. The dataset underwent pre-processed using correlation analysis and two different clustering algorithms: K-means and two-step for customer segmentation modeling. A pool of algorithms namely regression-based, decision trees and Neural Network algorithms were used to build our prediction model.

On the other hand, a predictive model presented by A. A. Bakar et. al. [8], on dengue outbreak had shown positive outcome by adopting multiple rule-based classifiers including Naïve Bayes, Decision Tree, Rough Set, and Associative Classifiers. The study was done using a set of data that consists of a total 8505 dengue patient records with 134 attributes involved. The prediction is conducted by first comparing the performance of each classifier. In the second stage, the study combined all classifiers and the results showed that this technique outperformed the performance of using single classifier.

One common feature in all the studies mentioned above is that the models' accuracy were evaluated by using true-positive rate and true-negative rate confusion matrix as their measurement, this is later adopted in this paper.

3 Methodology

3.1 Overview

In this section, the techniques used in this study will be explained in details. Comparisons of using different approaches including multiple model combination, i.e., EWD, K-means and Naïve Bayes are introduced to make predictions under a supervised environment.

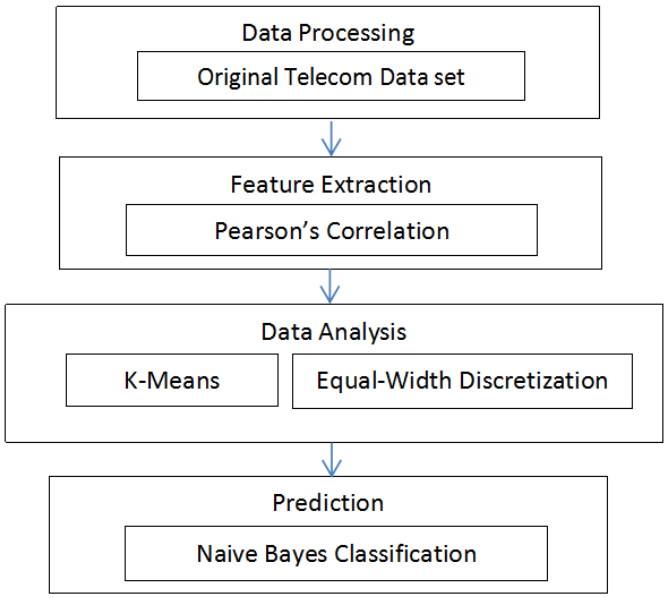

Fig. 1: Framework of prediction process

The following subsections will go through the entire process flow from data processing, feature selection and data analysis to the final validation of the model. The entire flow for the prediction model framework is presented in Figure 1.

3.2 Data Processing

The data in this study is obtained from SGI MLC++ package[1] that is originally on the UCI Machine Learning Repository. The dataset is a set of cleaned customer churn data from a telecommunications company. The features available are users' calling activity data along with churn label specifying the customer subscription. This dataset consists of 5,000 customer caller data and is randomly separated into 1,667 instances to be used as testing and model performance evaluation while the remaining 3,333 instances are kept for training and cross-validation purposes. These data are stored in CSV format and the list of attributes is shown in Table 1.

[1] ("Data Source Link http://www.sgi.com/tech/mlc/db/")

Table 1: Data type of attributes

Attributes	Type of attributes
State	Nominal
Account length	Discrete
Area code	Discrete
Phone number	Numerical
International plan	Boolean
Vmail plan	Boolean
Number vmail messages	Discrete
Total day minutes	Continuous
Total day calls	Discrete
Total day charge	Continuous
Total eve minutes	Continuous
Total eve calls	Discrete
Total eve charge	Continuous
Total night minutes	Continuous
Total night calls	Discrete
Total night charge	Continuous
Total intl minutes	Continuous
Total intl calls	Discrete
Total intl charge	Continuous
Num cust service calls	Discrete
Churn	Boolean

The dataset comprised of a variety of variable types, namely, nominal, continuous, discrete and Boolean. The unique key value of the dataset is the phone number of each user. While on the other hand, the prediction goal is to successfully classify the customer churn with only binary output; yes or no. Finally, 18 attributes are selected to carry out the prediction.

3.3 Feature Selection

Data preparation is a crucial phase before going into the data analysis. The main objective for preprocessing the data is to reduce the dimension of the dataset attributes and remove the unnecessary attributes using an appropriate method. A correlation analysis forms a filtered model that can effectively remove redundant attributes so that it will be less costly in computation. By using this approach, attributes that have a high correlation between each other will go through a series of filtration so that attributes that brings little meaning to the dataset will be eliminated. The dependency between two attributes is gauged using the commonly known ‘Pearson’s Correlation Coefficient’. The correlation coefficient equation is defined as below:

$$\begin{aligned} \mathrm{Corr}(X,Y) &= \frac{\mathrm{cov}(X,Y)}{\sigma_x \sigma_y} \\ &= \frac{E\left[(X-\mu_x)(Y-\mu_y)\right]}{\sigma_x \sigma_y} \end{aligned} \tag{1}$$

where μ_x and μ_y are the mean, while σ_x and σ_y are the standard deviations of X and Y respectively. E is the expected value operator, cov means covariance[10]. Attributes are considered to be highly correlated if the pair-wise correlation has +/-0.9 or higher, and one of the attributes can be omitted from the model.

3.4 Data Analysis

In this subsection, two novel methods (i) EWD combined with categorical Naïve Bayes classifier and, (ii) K-Means combined with categorical Naïve Bayes classifier, are discussed. Since Naïve Bayes classifier implemented in this study is targeted to classify qualitative data, also often called categorical data into a finite number of intervals. Thus, EWD and K-Means are introduced to transform those quantitative data into qualitative form. The data are transformed into a higher level of knowledge representation as opposed to the subtle individual values which helps to simplify the data mining and analysis process. This process is commonly known as data discretization, which is an indispensable step in the data analysis pipeline.

EWD algorithm is an unsupervised discretization technique where it takes no account of the class information when selecting the interval boundaries. When performing EWD algorithm, continuous data in the dataset is discretized into multiple number of bins consisting of equal range[11]. This algorithm involves sorting the observed values of the attributes in order to look for the maximum and minimum values within an attribute. The interval can be computed by dividing the range into user specified number of bins. Finally, the continuous data will be assigned to its corresponding interval/cluster.

On the other hand, K-Means carries out clustering process is also suitable to be used to transform the nominal and continuous data into categorical format. It is a non-hierarchical partitioning clustering process that calculates the continuous distance-based similarity measure to group the data. The values for each dataset are grouped together based on their Euclidean distance. The iterative K-Means aims to minimize the objective function by clustering n input data instances $x_1, x_2, \ldots, x_n$ into k disjoint clusters C_j, where $j = 1,2,3,\ldots k$. The objective function of K-Means namely squared error is shown below:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| X_i^{(j)} - c_j \right\|^2 \tag{2}$$

where $X_i^{(j)}$ represents the i-th data point in the cluster C_j and c_j is the centroid of cluster C_j. In the beginning, the number of clusters, k is predefined (randomly). The function will randomly select k points as the cluster centroid, c_j. For the next step, by computing the Euclidean distance between each data point and the centroid, the data point will continuously relocate to their closest cluster until no changes are found. The clustering process aims to minimize the sum of squared distance between data point within its own cluster over the sum of squared distance between data points from different clusters. Therefore, to obtain the optimal number of clusters, k for K-means, Elbow method is deployed for each attribute. The Elbow method identifies the number of clusters that achieves low Sum of Squared Error (SSE) as the values converge when they reach the minimum number of clusters.

For Naïve Bayes classifier in this study, the underlying technique is known as Bayes Theorem [12]. The technique has a preliminary assumption where every attribute, X from given class, c is said to be conditionally independent before it can be implemented in Naïve Bayes. This helps to simplify the operation of the algorithm. The algorithm will compute the posterior probability of the different combinations of attributes with the given class and return the predicted probability values. It will then predict the output by capturing the likelihood with the training set given. Below is the general expression of the Naïve Bayes posterior probability's formula:

$$P(c|X) = \frac{P(x_1 \cdots x_n|c) \cdot P(c)}{P(X)} \tag{3}$$

where c represents the given classes and X, is the predictors $x_1 \cdots x_n$ which correspond to the attributes within the dataset. $P(x_1 \cdots x_n|c)$ is the likelihood each instance of given class, c, and $P(c)$ is the prior probability of class.

3.5 Data Validation

Model evaluation process is presented using confusion matrix that computes the precision and recall of the results. The equation below estimates the overall performance of the model.

$$OA,\% = \frac{n_{cp}}{N} \times 100\% \tag{4}$$

This formula attains the overall accuracy (OA) of the model by finding the number of correctly predicted outputs, n_{cp} divided by the total number of samples, N. Below demonstrates the formula to compute the accuracy for the true positive and true negative categories.

$$\text{True Positive Rate} = \frac{\text{number of correctly predicted churn customers}}{\text{total number of churn customers}} \times 100\% \quad (5)$$

$$\text{True Negative Rate} = \frac{\text{number of correctly predicted non churn customers}}{\text{total number of non churn customers}} \times 100\% \quad (6)$$

4 Results and Discussions

The results for the two methods introduced in section 3 are discussed in the following subsections.

4.1 Analysis of Attributes

Figure 2 illustrates the correlation heat map matrix that is constructed based on the correlation of customer activities across 18 selected attributes. The correlation values are annotated as shown in the matrix. Highly correlated features will be indicated with darker shades; in contrast, light color represents low correlation between attributes. The results suggest that 5 pairs of attributes are highly correlated to each other. One of the correlated attributes in each pair can be removed from the prediction model, namely ‘voicemail message plan’, ‘total night charge’, ‘total day charge’, ‘total eve charge’ and ‘total intl charge’. The comparison of original dataset and reduced dataset is shown in Table 2.

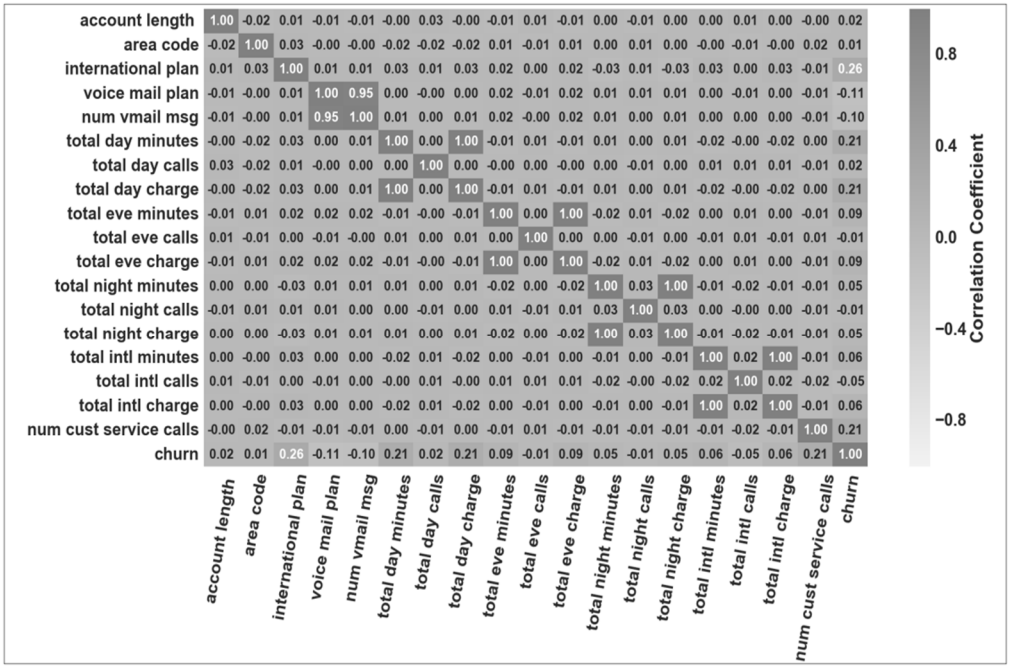

Fig. 2: Heatmap Correlation matrix plot

Table 2: Comparison of dataset attributes before and after selection based on correlation

Dataset	Number of attributes	Attributes
Original Dataset	18	'account length', 'area code', 'international plan', 'phone number', 'vmail plan', 'num vmail messages', 'total day minutes', 'total day calls, 'total day charge', 'total eve minutes', 'total eve calls, 'total eve charge', 'total night minutes', 'total night calls, 'total night charge', 'total intl minutes', 'total intl calls, 'total intl charge', 'num cust service calls'
Reduced Dataset	13	'account length', 'area code', 'international plan', 'vmail plan', 'total day minutes', 'total day calls, 'total eve minutes', 'total eve calls, 'total night minutes', 'total night calls, 'total intl minutes', 'total intl calls , 'num cust service calls'

In applying K-Means algorithm, the optimal number of clusters for each attribute in original dataset or reduced dataset is determined using Elbow method. One of the examples of getting the optimal number of clusters is shown in Figure 3.

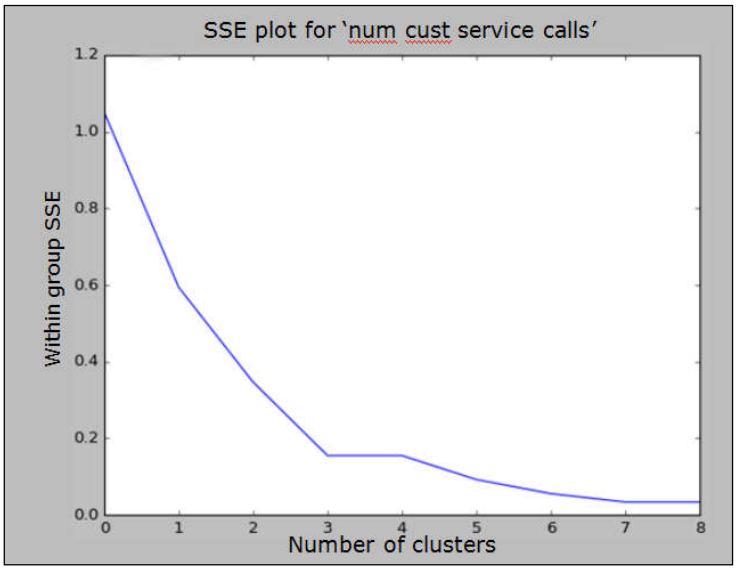

Fig. 3: K-Means clustering SSE plot to find number of clusters for 'num cust service calls'

Figure 3 illustrates the plot of SSE versus the number of clusters for the attribute 'num cust service calls'. It shows that the optimal number of clusters is best to set to three because the SSE decreases drastically when the number of clusters reaches three. For the remaining attributes, the optimal number of clusters is set to be four because at four the SSE decreases drastically when it reaches four. In the following section, the model that uses this configuration is named as Model A.

4.2 Simulation Results

Table 3 reflects the impacts of accuracy on the original dataset and reduced dataset. It is observed that by eliminating highly correlated attributes can effectively reduce the computational time by more than 30%. Next, a True-False analysis on reduced dataset, named as Model A is carried out and the results are shown in Table 4.

Table 3: Comparison of results between original and reduced dataset

Results	Original Dataset		Reduced Dataset	
	EWD	K-Means	EWD	K-Means
Overall accuracy	88.00%	86.50%	88.60%	86.72%
Computational time	391.4s	400.4s	276.9s	270.0s

In Table 4, the true positive value for K-Means combined Naïve Bayes classifier method shows slightly better compares to EWD combined Naïve Bayes classifier method. However, true positive values for both methods are below 50%. To solve this issue, a correlation analysis between the 13 attributes with churn results (as in Figure 2) is carried out and the results show the 'total day minutes' attribute is on e of the attributes that has the highest correlation value among non-Boolean type o f attributes.

Table 4: True-False analysis of Model A

EWD(%)		K-Means(%)	
TP	TN	TP	TN
42.41	93.41	43.75	88.14

Therefore, the model is fine-tuned by increasing the number of clusters for this attribute into 5, 6, 7, and 8, whereas the remaining attributes remain unchanged. These new models are named as B, C, D, and E as stated in Table 5.

Table 5: Description of models

Models	Description
B	'Total day minutes' divided into 5 clusters
C	'Total day minutes' divided into 6 clusters
D	'Total day minutes' divided into 7 clusters
E	'Total day minutes' divided into 8 clusters

A True-False analysis based on the Table 5 is carried out and the results are showed in Table 6.

Table 6: True –False analysis for Model B, C, D and E

Models	EWD(%)		K-Means(%)	
	TP	TN	TP	TN
B	25.89	96.10	43.75	90.56
C	43.75	88.69	54.46	77.66
D	59.82	66.71	62.29	60.66
E	79.91	46.53	62.29	57.77

The number of customers churned that is correctly detected using the K-Means algorithm is better compare to EWD in Model B, C and D. Among these three models, Model D has the highest accuracy of true positive value. For Model E, the true positive value is comparatively higher when EWD is implemented, but it has the lowest true negative value among all.

Table 7: Overall Accuracy of all models

Models	Overall Accuracy	
	EWD(%)	K-Means(%)
B	86.67	84.20
C	82.65	74.54
D	65.69	60.98
E	51.02	58.46

Table 7 presents the overall accuracy of the four models. The EWD models show better accuracy in terms of overall performance due to its higher sensitivity in predicting true negative instances. However, K-Means combined Naïve Bayes shows higher confidence in predicting potential churn customers. In short, Model

A, B, C and D have achieved overall accuracy above 60% and Model D has a better sensitivity to predict churn customers when the 'total day minutes' is divided to 7 clusters.

5 Conclusion and Future Works

In conclusion, the K-Means combined with Naïve Bayes method demonstrated better sensitivity and accuracy for predicting customer churn in telecommunications sector. The results obtained after optimization of parameters for Model A has proven that by increasing the number of clusters of the corresponding attribute helps in improving the true positive performance of the model. However, this technique experiences one shortcoming which is the trade-off point in correctly predicting the true positive and true negative output. The result obtained has addressed the impact of class imbalance problem which makes it difficult for classifier to make prediction. The training set that consists of 3,333 instances has only 14.49% of churned customers while the remaining 85.50% is non-churn instances. Thus, this requires further investigation theoretically and experimentally by considering several pertinent issues. One of the approaches to improve the model is to study the interval boundaries during data discretization process; also, the number of intervals could possibly be another factor that affects the learning rate of the classifier. Lastly, this technique can also be improved by adopting different machine learning algorithm such as support vector machine, decision tree as well as the Bayesian network that allows learning of non-linear data sample.

ACKNOWLEDGEMENTS.
This work was funded by the TM Research & Development (TM R&D) Grant.

References

[1] I. Khan, I. Usman, T. Usman, G. U. Rehman, and and Ateeq Ur Rehman. (2013). Intelligent Churn prediction for Telecommunication Industry, *Int. J. Innov. Appl. Stud.*, 4(1), 165–170.

[2] G. M. Weiss. (2009). Data Mining in the Telecommunications Industry, *Data Mining and Knowledge Discovery Handbook*., 1189-1201.

[3] C. P. Wei and I. T. Chiu. (2002). Turning telecommunications call details to churn prediction: A data mining approach, *Expert System with Applications,* Vol. 23(2), 103–112.

[4] C.cary. (2015). Telcos must turn big data into smart data to manage customer churn and loyalty. [https://www.ovum.com/press_releases/telcos-must-turn-big-data-into-smart-data-to-manage-customer-churn-and-loyalty].

[5] L. Pandeeswari and K. Rajeswari. (2015) "K-Means Clustering and Naive Bayes Classifier For Categorization Of Diabetes Patients," *IJISET - Int. J. Innov. Sci. Eng. Technol.*, 2(1), 179–185.

[6] G. Vennila, N. S. Shalini, and M. S. K. Manikan. (2015). Navie bayes intrusion classification stem for VoIP network using honeypot, *Int. J. Eng. Trans. A Basics*, 28(1), 46–53.

[7] A. Rehman and A. R. Ali. (2014). Customer Churn Prediction , Segmentation and Fraud Detection in Telecommunications Industry*, BioMedCom 2014 Conference on Biomedical* (pp. 1-9).

[8] A. A. Bakar, Z. Kefli, S. Abdullah, and M. Sahani. (2011). Predictive models for dengue outbreak using multiple rulebase classifiers, *Proc. 2011 Int. Conf. Electr. Eng. Informatics*, ICEEI.

[9] Y. Yang and G. I. Webb. (2009). "Discretization for naive-Bayes learning : managing discretization bias and variance", *Machine Learning*, 74(1), 39-74.

[10] S. Munirathinam and B. Ramadoss. (2016). "Predictive Models for Equipment Fault Detection in the Semiconductor Manufacturing Process," *Int. J. Eng. Technol.*, 8(4), 273–285.

[11] R. Dash, R. L. Paramguru, and R. Dash. (2011). Comparative Analysis of Supervised and Unsupervised Discretization Techniques, *Int. J. Adv. Sci. Technol.*, 2(3), 29–37.

[12] H.-C. Kim and Z. Ghahramani. (2012). Bayesian Classifier Combination, *Proc. Int. Conf. ArtificalIntell. Stat. on* ,Vol. 11, 619–627.

[13] Zanariah Zainudin and Siti Mariyam Shamsuddin. (2016). Predictive Analytics in Malaysian Dengue Data from 2010 until 2015 using BigML, *International Journal of Advances in Soft Computing and Its Applications,* Vol. 8(3), 18-30.