Credit Card Fraud Detection Using RUS and MRN Algorithms

Seminar Report

Submitted by

ADHIL AHAMED A.P

IDK16CS002

to

the APJ Abdul Kalam Technological University
in partial fulfillment of the requirements for the award of the degree

of

BACHELOR OF TECHNOLOGY

in

Computer Science and Engineering

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

GOVERNMENT ENGINEERING COLLEGE IDUKKI

PAINAVU-685603

November 2019

GOVERNMENT ENGINEERING COLLEGE
PAINAVU IDUKKI-685 603

CERTIFICATE

This is to certify that the seminar report entitled **"Credit Card Fraud Detection Using RUS and MRN Algorithms"** *has been submitted by* **ADHIL AHAMED A.P (IDK16CS002)** *in the partial fulfillment for the award of B.Tech Degree in COMPUTER SCIENCE AND ENGINEERING for the academic session 2016-2020 under the supervision and guidance of Ms. Liya Joseph, Department of Computer Science and Engineering,Govt. Engineering College, Idukki.*

Seminar Guide
Ms. Liya Joseph
Assistant Professor
Dept. of CSE
GEC Idukki

Seminar Coordinator
Ms. Deepa S S
Associate Professor
Dept. of CSE
GEC Idukki

Head of the department
Dr. Madhu KP
Associate Professor
Dept. of CSE
GEC Idukki

ACKNOWLEDGEMENT

I give all honour and praise to the **GOD** who gave me wisdom and enabled me to complete my seminar on" Credit Card Fraud Detection Using RUS and MRN Algorithms" successfully.

I express my sincere thanks to **Dr. Satheesh Kumar, Principal, Government Engineering College, Idukki,** for providing the right ambiance to work on the seminar.

I would like to extend my sincere gratitude to **Dr. Madhu KP, Head of Department, Computer Science and Engineering** for permitting me to work on the seminar and for his guidance, encouragement, support and care throughout the entire period of my course of study.

I deeply indebted to my Seminar Coordinator **Ms. Deepa S S, Associate Professor, Department of Computer Science and Engineering** for her continued support throughout our seminar.

It is with great pleasure that I express my deep sense of gratitude to my seminar guide **Ms. Liya Joseph, Assistant Professor, Department of Computer Science and Engineering** for her guidance, supervision, encouragement and valuable advice in each and every phase of my seminar.

I would like to thank all other faculty members and the fellow students of Government Engineering College, Idukki, for their warm friendship, support and help.

Also, I express my hearty thanks to my beloved parents for their love, encouragement and dedication in shaping my career.

ADHIL AHAMED A.P

ABSTRACT

Today's enterprises increasingly use credit cards as the major method of payment for products and services. This method provides faster and easier payments, but also introduces a possibility of credit card frauds which is a major issue that needs to be addressed.

The proposed model presents an effective and efficient way of analyzing patterns in payment of credit card users to detecting credit card fraud. Here MRN algorithms, which consist of Multi-Layer Perceptron, Radial Basis Function and Naïve bayes classifiers, are used along with a RUS sampling technique that selects the training data sets from an enormous dataset to classify payments as fraud or not (RUSMRN). Using customer transactions and repayments, financial information and statements as guidelines, the system predict risk of customer business operations.

The proposed system improves classification accuracy of unbalanced characteristic data. By doing so, we improve the safety of credit card usage by denying the identified transactions and putting them up for review.

Contents

List of Figures

Chapter 1

INTRODUCTION

In today's world, credit cards are becoming increasingly used as the preferred method of payment for all kinds of products and services. This increase in preference of credit cards is because of them being faster in terms of providing with credit and easier payments facilitated by the systems put as an alternative for payment by cash.

It has been evidenced that this use of credit cards is a double edged sword. The ease provided by the non-requirement of an authentication, such as the OTP in case of debit cards or payment in cash, also gives way to the possibility of credit card frauds the premise being that only the credit card or the number by itself is enough for making payments. This invites frauds by various people who can come by the possession of the card or the number. Such a feature makes the financial system fragile since credit card frauds accounts for significant loses of all the loses in the system presenting a crisis for credit card companies and individual using them. As frauds of these kind increases, the credibility of credit card payments and trust in these companies decreases.

A good management of credit cards is to be put in place which can detect and prevent frauds. The management applies financial information, financial statements, customer transactions and repayments as guidelines to

enforce detection and prevention of credit card frauds.

Application of machine learning is one way of doing so. The proposed system uses an ensemble of some ML techniques, namely an RUSMRN ensemble, for detection of credit card fraud. It applies data sampling technique called as RUS for adjusting the class distribution of the training data set and thus improves the problem of class imbalance.

Chapter 2

RELATED WORK

2.1 Data mining techniques for the predictive default

This research[1] aimed at the case of customers' default payments in Taiwan and compares the predictive accuracy of probability of default among six data mining methods. From the perspective of risk management, the result of predictive accuracy of the estimated probability of default will be more valuable than the binary result of classification - credible or not credible clients. Because the real probability of default is unknown, this study presented the novel ''Sorting Smoothing Method" to estimate the real probability of default. With the real probability of default as the response variable (Y), and the predictive probability of default as the independent variable (X), the simple linear regression result (Y = A + BX) shows that the forecasting model produced by artificial neural network has the highest coefficient of determination; its regression intercept (A) is close to zero, and regression coefficient (B) to one. Therefore, among the six data mining techniques, artificial neural network is the only one that can accurately estimate the real probability ofdefault.

In recent years, the credit card issuers in Taiwan faced the cash and credit card debt crisis and the delinquency is expected to peak in the third quarter of 2006 (Chou, 2006). In order to increase market share, card-issuing banks in Taiwan over-issued cash and credit cards to unqualified applicants. At the same time, most cardholders, irrespective of their repayment ability, overused credit card for consumption and accumulated heavy credit and cash– card debts. The crisis caused the blow to consumer finance confidence and it is a big challenge for both banks and cardholders.

2.2 Prevention of Credit Card Fraud Detection based on HSVM

Specific crime in the banking system is credit card fraud. Credit card usage has been increased due to the rapid growth of E-commerce techniques. Credit card fraud also increased at the same time. Prevention is better than detection. So the existing system prevented the credit card fraud by identifying fraud in the application of the Credit card. Due to the limitation of the existing system, this paper proposed new algorithm along with the existing algorithm. Scalability issues, extreme imbalanced class and time constraints are the limitation of existing systems. Those limitations are overcome by hybrid support vector machine (HSVM)[9] along with communal and spike detection for credit card application fraud detection. HSVM is the most used method for the pattern recognition and classification.

Solving business problems in banking sectors are done mostly by contribution of Data Mining. Better targeting, acquiring new customers, and fraud detection in credit cards, fraudulent transactions can be done by Data mining techniques. Data Mining is one of the most promising interdisciplinary developments in Information Technology. The main target that focused on this system is to preserve the credit fraud in the initial stage of the credit life cycle. The implementation of this algorithm in order to perform

the identification of frauds, this system uses the hybrid support vector machine (HSVM) for computing the weight of the each attribute for communal and spike detection for credit card application fraud detection. . This project has been proposed with the efficiency in scalability by updating the evaluation of data.

2.3 Real-time Credit Card Fraud Detection Using Machine Learning

Credit card fraud events take place frequently and then result in huge financial losses . The number of online transactions has grown in large quantities and online credit card transactions holds a huge share of these transactions. Therefore, banks and financial institutions offer credit card fraud detection applications much value and demand. Fraudulent transactions can occur in various ways and can be put into different categories. This paper[10] focuses on four main fraud occasions in real-world transactions. Each fraud is addressed using a series of machine learning models and the best method is selected via an evaluation. This evaluation provides a comprehensive guide to selecting an optimal algorithm with respect to the type of the frauds and we illustrate the evaluation with an appropriate performance measure. Another major key area that we address in our project is real-time credit card fraud detection. For this, we take the use of predictive analytics done by the implemented machine learning models and an API module to decide if a particular transaction is genuine or fraudulent. We also assess a novel strategy that effectively addresses the skewed distribution of data. The data used in our experiments come from a financial institution according to a confidential disclosure agreement.

Credit card fraud detection has been a keen area of research for the researchers for years and will be an intriguing area of research in the coming future. This happens majorly due to continuous change of patterns in frauds. In this paper, we propose a novel credit-card fraud detection system by detecting four different patterns of fraudulent transactions using best suiting algorithms and by addressing the related problems

identified by past researchers in credit card fraud detection. By addressing real time credit-card fraud detection by using predictive analytics and an API module the end user is notified over the GUI the second a fraudulent transaction is taken place. This part of our system can allow the fraud investigation team to make their decision to move to the next step as soon as a suspicious transaction is detected. Optimal algorithms that address four main types of frauds were selected through literature, experimenting and parameter tuning as shown in the methodology. We also assess sampling methods that effectively address the skewed distribution of data. Therefore, we can conclude that there is a major impact of using resampling techniques for obtaining a comparatively higher performance from the classifier. The machine learning models that captured the four fraud patterns (Risky MCC, Unknown web address, ISO Response Code, Transaction above 100$) with the highest accuracy rates are LR, NB, LR and SVM. Further the models indicated 74%, 83%, 72% and 91% accuracy rates respectively. As the developed machine learning models present an average level of accuracy, we hope to focus on improving the prediction levels to acquire a better prediction. Also, the future extensions aim to focus on location-based frauds.

2.4 Credit Card Fraud Detection through Parenclitic Network Analysis

The detection of frauds in credit card transactions is a major topic in financial research, of profound economic implications. While this has hitherto been tackled through data analysis techniques, the resemblances between this and other problems, like the design of recommendation systems and of diagnostic/prognostic medical tools, suggest that a complex network approach may yield important benefits. In this paper[11] we present a first hybrid data mining/complex network classification algorithm, able to detect illegal instances in a real card transaction data set. It is based on a recently proposed network reconstruction algorithm that allows creating representations of the deviation of one instance from a reference group. We show how the inclusion of features

extracted from the network data representation improves the score obtained by a standard, neural network-based classification algorithm and additionally how this combined approach can outperform a commercial fraud detection system in specific operation niches. Beyond these specific results, this contribution represents a new example on how complex networks and data mining can be integrated as complementary tools, with the former providing a view to data beyond the capabilities of the latter.

Complex networks and data mining models share more characteristics than what we could have expected in the first naive approach, most notably having similar objectives: both aim at extracting information from (potentially complex) systems to ultimately generate new compact quantifiable representations. At the same time, they approach this common problem from two different approaches: the former by extracting and quantitatively evaluating the underlying structure; the latter by creating predictive models based on historical data. In this paper we test the hypothesis that complex networks can be used as a way to improve data mining models, framed within the problem of detecting fraud instances in credit card transactions, providing a new example about how complex networks and data mining may be integrated as complementary tools in a synergistic manner in order to improve the classification rates obtained by classical data mining algorithms.

Results confirm that features extracted from a network-based representation of data, leveraging on a recently proposed parenclitic approach, can play an important role: while not effective in themselves, such features can improve the score obtained by a standard ANN classification model. We further show how the resulting model is especially efficient in detecting frauds in some niches of operations, like medium-sized and on-line transactions. Finally, we illustrate as in the latter case that the network-based model is able to yield better results than a commercial fraud detection system. All results have been obtained with a unique data set, comprising all transactions managed during two years by a major Spanish bank and including more than million operations.

Chapter 3

PROPOSED SYSTEM

Implementation of this system uses RUSMRN algorithms which consist of sampling techniques and ML classification algorithms. It is based on linear mapping, non-linear mapping and probability theory with RUS for classifying class imbalance problems. RUSMRN is based on AdaBoost algorithm used to combine different models to give a stronger classifier[2]. The data sampling technique RUS is applied to solve the class imbalance problems. The system uses customer transactions, repayments, financial information and statements as input guidelines for classifying records.

3.1 RUSMRN

In this paper we propose RUSMRN ensemble for classifying default of the data sets for client credit cards. RUSMRN is an ensemble model that stands for Random Under-Sampling and MLP, RBF, Naïve Bayes classifiers which are based on linear mapping, non-linear mapping and probability theory. RUS is a sampling technique which is used to solve the problem of class imbalance by adjusting the class distribution of the training data set. MRN applies a boosting algorithm, AdaBoost, which is similar to a decision tree that applies weights to predict the response of input.

3.2 Problem of class imbalance

Any learning algorithm or model is only as good in terms of performance as the quality and quantity of the dataset it is provided with for learning. In cases similar to ours, there exists a situation where the class distribution of the dataset is imbalanced. This means that, in a dataset with two classes, say positive and negative class, the number of records belonging to one class is very much larger than the number of records belonging to the other class.

The ideal case is that the number of records of both classes is roughly equal. Thus it is not preferred that number of records of one class is far less than the number of records of the other class. In such a case, the model tends to classify the records belonging to minority class as a record in the majority class. This problem of class imbalance is seen in various situations like fraud detection, anomaly detection, medical diagnosis, oil spillage detection, facial recognition, etc. This can be explained by the fact that the model needs more data to learn from, specifically to learn to identify negative class records[7].

3.3 Random Under-Sampling

Random Under-sampling is a data sampling technique which adjusts the class distribution of the training data set , alleviating the problem of class imbalance[8]. This is one of the most commonly used data sampling technique. There are two methods for data sampling, Undersampling and oversampling. Undersampling is

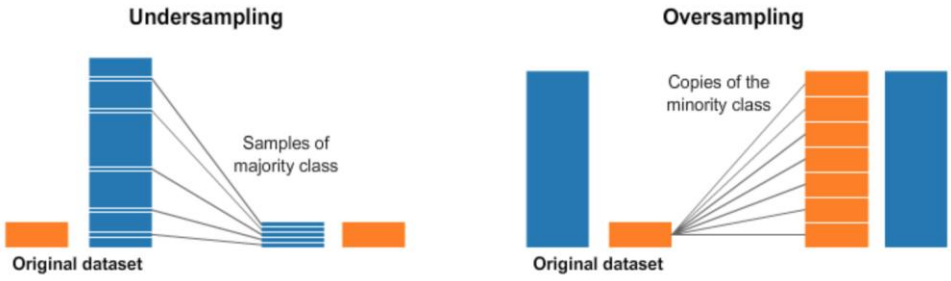

Figure3.1: Examples of undersamplig and oversampling

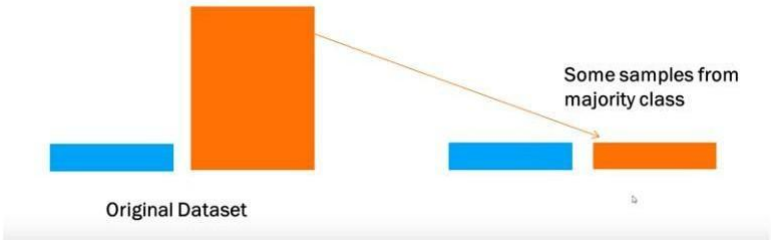

Figure 3.2: Example of RUS

the removal of examples from the majority class to achieve desired class distribution. Oversampling is the adding of examples to the minority class till desired class distribution is achieved.

Here RUS is used which applies undersampling by simply removing examples in the majority class at random until a desired class distribution is achieved, un-like other undersampling techniques where multiple copies of some of the minor-ity classes are used. Although it alleviates class imbalance problem, it increases the variance of the classifier and sometimes may remove potentially useful or important samples. Studies have shown that combination of undersampling with ensemble learning can be done to achieve better performance.

Take an example where there is a dataset with a total of 1000 records with only 100 records account for negative class. When RUS is applied here, 100 records are selected randomly from the positive class and is combined with the 100 records in the negative class to get a total of 200 records. This solved the class imbalance problem and is viewed as a dataset with only 200 records. But as compared to the original dataset with 1000 records, the undersampled one is very small, thus decreases the learning dataset size.

3.4 MRN Algorithm

MRN algorithm is an ensemble algorithm that stands for MLP (Multi Layer Perceptron), RBF (Radial Basis Function) and Naïve bayes classifier algorithm. It a new hybrid ensemble model which is based on three learning models, Linear mapping, Non-linear mapping and Probability theory, which is implemented by MLP, RBF and Naïve bayes classifier algorithm respectively. The three algorithms are used to create an ensemble using a boosting technique, namely AdaBoost, that combines these three learning models and gives a stronger classifier that overcomes shortcomings and applies advantages of the individual learning models. Here this is done by AdaBoost that is a weight based technique which applied these weights to each model to predict the response or output of the data[12].

3.5 MLP network

A multi layer perceprton is an Artificial Neural Network based on the percep-tron model, the main difference being the number of hidden layers present in the model. It is a feedforward ANN model, that is, the flow of data is only in the for-ward direction, no error is propagated back to the layer before. It is the foundation of deep learning models and is commonly used for supervised learning.

MLP classifies the records as negative or positive class by mapping the sets of input data onto a set of appropriate outputs or classes. The output is the based on the activation function applied in each node combined with weights of each connections. The activation function used here is the sigmoid function. Sigmoid is preferred since it gives an output in the range of 0 to 1 which is suitable in this application.

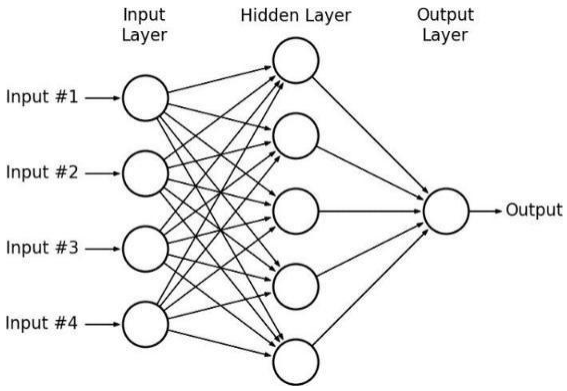

Figure 3.3: A hypothetical example of Multilayer-perceptron network

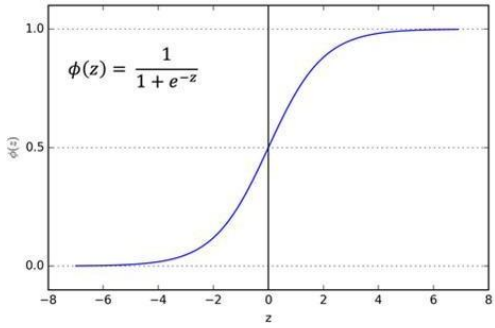

Figure 3.4: Sigmoid Activation function

3.6 RBF Network

Radial Basis Function network is a type of ANN that uses RBF as its activation function. Radial basis function is a function which works similar to the K-nearest Neighbor model. The neurons in hidden layer apply RBF which consist of gaussian functions that centered on a point. Similar to MLP, this also applies weighted sums in each layer.

RBF network works on the basis that predicted value of an item is close to other items with same predictor variables. Centers are chosen whose class label is

already known randomly and hypotheses made based on these centers. Hypothesis, h(x), depends on these centers and the standard deviation of items from the centers defined in each RBF unit. The network may differ in terms of number of hidden layers as it is determined by the number of centers selected in the RBF units. It is a more intuitive approach since it measures the input's similarities with examples from the training data set. The standard deviation can be either Euiclidean or Manhattan distance.

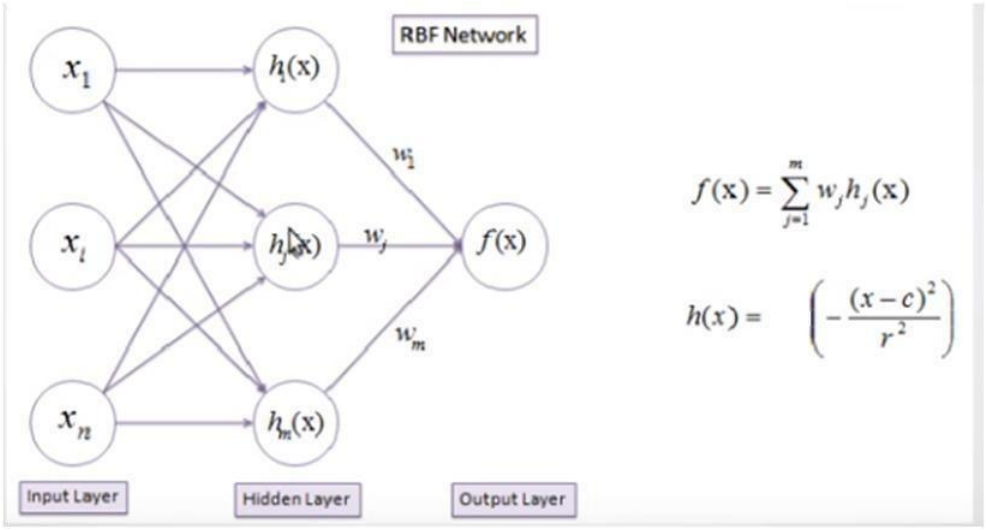

Figure 3.5: RBF Network, Function and Hypothesis

3.7 Naïve Bayes Classifier

Naïve Bayes classifier is a classification model based on the Bayes theorem. It uses the probability and conditional probabilities for classifying unseen records into classes for each class. It assigns probability values for each feature value for each instance to be classified. The total probability of an instance belonging to a class is calculated as follows:

$$P(C_i|X) = \frac{P(C_i)\prod_{k=1}^{n} P(X_k|C_i)}{P(X)} \quad (1)$$

where Ci is the class label and X is an instance of unseen data to be classified.

This works well for certain nearly-functional feature dependencies, thus reaching its best performance in two opposite cases: completely independent features (as expected) and functionally dependent features (which is surprising).It is surprisingly effective in practice since its classification decision may often be correct even if its probability estimates are inaccurate.

3.8 AdaBoost Algorithm

AdaBoost stands for Adaptive Boosting which is a machine learning meta-algorithm. This adds up different learning models to create a strong classifier. It can be used in conjunction with many other types of learning algorithms to improve performance. Each model is fitted on modified versions of the original dataset.

Error is evaluated for each model based on the training dataset. Initially, each model is given equal weights. These weights are adjusted in each iteration based on the error evaluated each time. Higher weights are given to models with lower errors and lower weights are given to models with higher errors. Also weights are increased for each incorrectly classified observation and weights are decreased for each correctly classified observation to increase the focus in the next iteration.

AdaBoost is said to be adaptive since subsequent weak learners are tweaked in favor of those instances misclassified by previous classifiers. The output of the different learning algorithms is combined into a weighted sum that represents the final output of the boosted classifier which can be proven to converge to a strong learner.

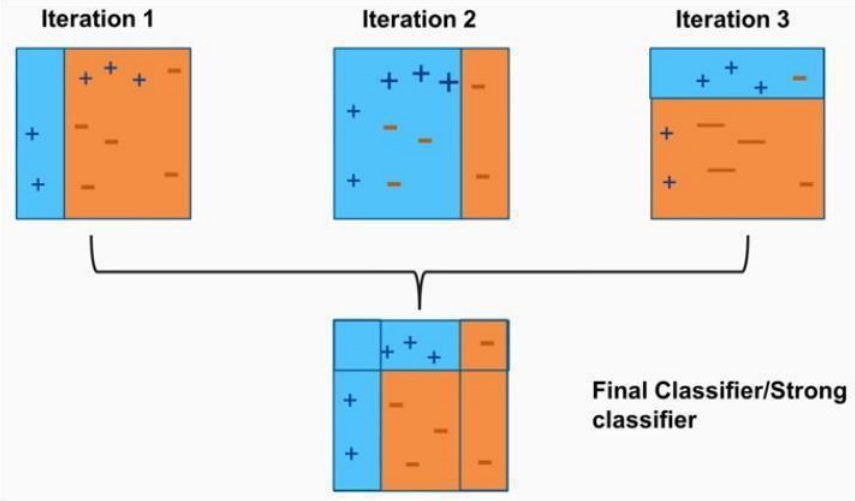

Figure 3.6: AdaBoost Example

Algorithm: An Ensemble Model based on linear mapping, non-linear mapping and probability for classification with class imbalance problem

Input: Training data S={x_1,x_2,….x_n}, $x_i \in X$ with correct class $y_i \in \Omega$, Ω={y_1,y_2,…y_C}

Output: $$H(x) = \arg max_{y\in\Omega} \sum_{t=1}^{T} \ln\left(\frac{1}{\beta_t}\right)[h_t(x) = y] \quad (2)$$

Method:

1. Initialize distribution D(i)=1/n ,i=1,2,…,n
2. Create temporary training data S_t' with distribution D_t' by applying random undersampling.
3. Train MLP weak learner with training set to receive hypothesis h_a:X→Ω
4. Compute the error of h_a: $\varepsilon_a = \sum_{i=1}^{n} I[h_a(x_i) \neq y_i].D_t(i)$ (3)
   where ε_a is the error of h_a and y_i is the desired class.
5. If ε_a>0.5 then go back to step 2.
6. Calculate the normalized error $\beta_a = \varepsilon_a/(1-\varepsilon_a)$ (4)
7. Update the distribution D_a: $D_{a+1}(i) = \frac{D_t(i)}{z_a} \times \begin{cases} \beta_a \; if \; h_a(x) = y_i \\ 1 \quad Otherwise \end{cases}$ (5)
   where z_a is a normalization constant.
8. Train weak learner RBF with training set to receive hypothesis h_b:X→Ω
9. Compute the error of h_a: $\varepsilon_b = \sum_{i=1}^{n} I[h_b(x_i) \neq y_i].D_t(i)$ (6)
   where ε_b is the error of h_b and y_i is the desired class.
10. If ε_b>0.5 then go back to step 7.
11. Calculate the normalized error $\beta_b = \varepsilon_b/(1-\varepsilon_b)$ (7)

12. Update the distribution D_b: $D_{b+1}(i) = \frac{D_t(i)}{z_b} \times \begin{cases} \beta_b \ if \ h_b(x) = y_i \\ 1 \quad Otherwise \end{cases}$ (8)
    where z_b is a normalization constant.
13. Train weak learner Naïve Bayes with training set to receive hypothesis $h_c: X \rightarrow \Omega$
14. Compute the error of h_c: $\varepsilon_c = \sum_{i=1}^{n} I[h_c(x_i) \neq y_i].D_t(i)$ (9)
    where ε_c is the error of h_c and y_i is the desired class.
15. If $\varepsilon_c > 0.5$ then go back to step 12.
16. Calculate the normalized error $\beta_c = \varepsilon_c/(1 - \varepsilon_c)$ (10)
17. Update the distribution D_c: $D_{c+1}(i) = \frac{D_t(i)}{z_c} \times \begin{cases} \beta_c \ if \ h_c(x) = y_i \\ 1 \quad Otherwise \end{cases}$ (11)
    where z_c is a normalization constant.

Let x_i be a point in training data S and y_i is a class label in a set of class labels Ω.

In step 1, the weights of each instances in the training set are initialized to 1/n where n is the number of instances in the training data set. In step 2, RUS is applied to remove the majority class examples until 35% of the new training data S_t' . In step 3, train the base classifier MLP and obtain its training error rate which creates the weak hypothesis h_a. In step 4, compute the error ε_a of h_a. In step 6, calculate the normalized error beta a is calculated as $\varepsilon a/(1 - \varepsilon_a)$. Next, the weight distribution for the next classifier D_{a+1} is updated in steps 7. In step 8-16, RUSMRN train base classifiers RBF and Naïve Bayes by new dataset after that update the weight distribution for the next classifier by using the normalized error β_a. The final step, for the incoming unseen data, t calculated the class label of them by the following equation.

$$H(x) = \arg max_{y \in \Omega} \left\{ \ln\left(\frac{1}{\beta_a}\right)[h_a(x) = y] + \ln\left(\frac{1}{\beta_b}\right)[h_b(x) = y] + \ln\left(\frac{1}{\beta_c}\right)[h_c(x) = y] \right\} \quad (12)$$

Where H(x) the final hypothesis and h is the hypothesis of each classifier.

Chapter 4

EXPERIMENTAL RESULTS

For this experiment, the machine learning methods described in section 3 are trained to predict the dataset. Four classifiers method are used in this experiment consisting of the propose RUSMRN, RUSBoost, AdaBoost and Naïve Bayes clas-sifiers. The experiments are constructed on Matlab.

4.1 Dataset

The payment data of October, 2005, issued by an important bank from Taiwan is taken as the dataset and the targets are the credit card holders of the bank[1]. The total number of observations in the dataset is about 25,000 observations of which 5,529 observations are the cardholders with default payment from UCI machine learning database, that is, the negative class . This dataset employed a binary variable – default payment for class Yes represented by 1 and class No represented by 0 as the response variable. The column of 23 variables can be explained as:

- D1: Number of the given credit card: it includes both the individual consumer credit.

- D2: Gender (1 = male; 2 = female).
- D3: Education (1 = graduate school; 2 = university; 3 = high school; 4= others).
- D4: Marital status (1 = married; 2 = single; 3 = others).
- D5: Age (year).
- D6–D11: History of payment.
- D12 = Amount of bill statement in September, 2005;
- D13 = Amount of bill statement in August, 2005;
- D17 = Amount of bill statement in April, 2005.
- D18 = Amount paid in September, 2005.
- D19 = Amount paid in August, 2005.
- D23 = Amount paid in April, 2005.
- D19 = Amount paid in August, 2005.
- D23 = Amount paid in April, 2005.

4.2 Performance Measure

In this paper, the performance of the proposed system is measured by sensitivity, specificity and accuracy described as follows:

- Accuracy (ACC) is the overall success rate of the classifier defined as

$$ACC = (TP+TN) / (P + N) \quad (13)$$

where TP is the true value of positive rate, P is the positive class or yes class, N is the negative class or no class.

- Sensitivity or the true positive rate (TPR) which is defined as the fraction of positive instances predicted correctly by the model defined as

$$\text{Sensitivity} = TP / (TP+FN) \quad (14)$$

- Specificity is the true negative rate (TNR) which is defined as the fraction of negative instances predicted correctly by the model defined as

$$\text{Specificity} = TN / (FP+TN) \quad (15)$$

4.3 Results

The results of four model techniques are compared in the experiments. All model techniques are trained and three fold cross validation is used to investigated them. In the experiments, each technique are run three times and average results are shown in Figure.4.1

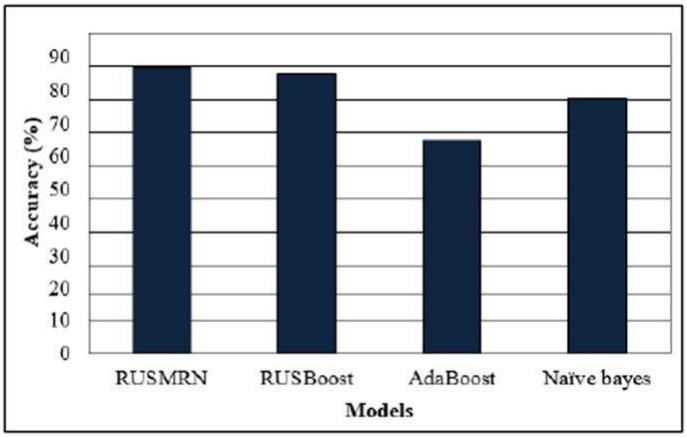

Figure 4.1: Average Accuracy of Four Classifiers

Figure 4.1 shows the average accuracy of four classifiers on three fold cross-validation. From the experimental results, it can be seen that the proposed RUSMRN model gives the highest accuracy than the others with 79.73 % while RUSBoost, AdaBoost and Naïve bayes can produce the average accuracy of 77.8%, 57.73% and 70.13% respectively. The ac- curacy of each class is also important because if the classifier predicts incorrectly,it may be a detriment to the customer. Therefore, the sensitivity and specificity value is applied in the experiments for evaluating the erformance of the proposed methods.

4.4 True Positive

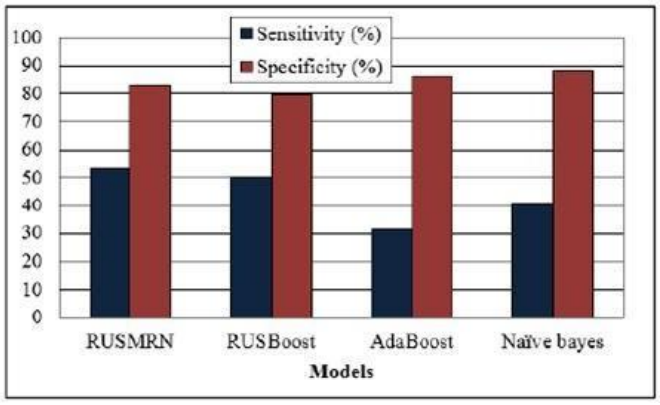

Figure 4.2: Comparison of Sensitivity and Specificity Four Classifiers

Figure 4.2 illustrates the averages of sensitivity and specificity run on three time. The sensitivity of the proposed RUSMRN is slightly higher than other methods at 53.36% and the sensitivity of RUSBoost, Ad- aBoost and Naïve bayes are in 50.3%, 31.4% and 40.% respectively. For specificity value, the specificity of Naïve bayes is higher than other methods at 88.13% and the specificity of

AdaBoost, RUSMRN and RUSBoost are in 86.16%, 83.1% and 79.8% respectively. According to the experimental results, it can be concluded that RUSMRN is appropriated for predicting the data because it has maximum of sensitivity and high specificity as a result with this highest accuracy model.

Chapter 5

CONCLUSION

In this paper, we present the predictive model by using machine learning methods that is called RUSMRN algorithm. Moreover, it is based on RUS data sampling technique and MRN algorithm to predict the data payment of October, 2005 issued by an important bank from Taiwan. The proposed RUSMRN algorithm combines boosting and data sampling to improve classification accuracy of unbalance characteristic data. From the experimental results, it can be seen that RUSMRN classifier outperforms the other procedures in terms of accuracy. In addition, it has highest sensitivity after training and testing by applying the pro-pose method. it is appropriated for predicting the data because it has the best classification performance in terms of accuracy and sensitivity.

Three fold cross validation done on each model and it is found that this model is the best suited one over the class imbalance problem. It also gives the highest accuracy than the other models with 89.73 %. It is the most appro- priate model for this problem of credit card fraud detection since it has maximum sensitivity and high specificity. But still, in the application, even the smallest percent of misclassification is undesirable.

REFERENCES

[1] I-Cheng Yeh and Che-hui Lien, "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients," Expert Systems with Applications, 36 (2009) 2473–2480

[2] A. Li1, W. Li1 and Y. Shi, "Study on the Application of Data Mining Algo-rithms in Credit Card Management," International Conference on EBusiness and Information System Security, IEEE, pp. 1 - 5, May 2009.

[3] R. YANG, X. ZHOU and W. Wang, "Is the small and medium-sized enterprises'credit default behavior affected by their owners' credit features?," IEEE , Management and Service Science, pp. 1 - 4, Aug. 2011.

[4] H. Yeh, M. Yang and L. Lee , "An Empirical Study of Credit Scoring Model for Credit Card," Innovative Computing, Information and Control, IEEE, pp. 216 - 219, Sept. 2007.

[5] W. Li and J. Liao, "An Empirical Study on Credit Scoring Model for Credit Card by using Data Mining Technology," Computational Intelligence and Security, IEEE, pp. 1279 - 1282, Dec. 2011

[6] S. S. Alkhasov, A. N. Tselykh and A. A. Tselykh, "Application of Clus-ter Analysis for the Assessment of the Share of Fraud Victims among Bank

Card Holders," International Conference on Security of Information and Net-works, ACM, pp. 103-106, 2015.

[7] C. Seiffert, T. M. Khoshgoftaar, J. V. Hulse and A. Napolitano, "RUSBoost: A Hybrid Approach to Alleviating Class Imbalance," IEEE Transl. Systems, Man, and Cybernetics, vol. 40, pp. 185 - 197, January 2010.

[8] Charleonnan and S. Jaiyen, "A new ensemble model based on linear mapping, nonlinear mapping, and probability theory for classification problems," International Joint Conference onComputer Science and Software Engineering, IEEE, pp. 88 - 92, July 2015.

[9] V.Mareeswari and Dr G. Gunasekaran, "Prevention of credit card fraud detection based on HSVM", International Conference On Information Communication And Embedded System(ICICES 2016), 978-1-5090-2552-7

[10] Anuruddha Thennakoon, Chee Bhagyani, Sasitha Premadasa, Shalitha Mihiranga and Nuwan Kuruwitaarachchi, "Real-time Credit Card Fraud Detection Using Machine Learning", INSPEC Accession Number: 18868933

[11] Massimiliano Zanin, Miguel Romance, Santiago Moral, and Regino Criado, "Credit Card Fraud Detection through Parenclitic Network Analysis", Complexity Volume 2018, Article ID 5764370, 9 pages

[12] A. Asuncion and D. J. Newman. (2007). UCI Machine Learning Repository [Online]. Available: http://www.ics.uci.edu/mlearn/ MLRepository.html