A PREDICTION OF CUSTOMER RELATIONSHIP MANAGEMENT MODEL IN BANKING SECTOR USING CHI-SQUARE AND SVM-RBF

Kayode O. Alabi, Jumoke F. Ajao, Suleiman O. Abdulsalam , Micheal O. Arowolo, Kareem A. Gbolagade

Department of Computer Science, Kwara State University, Malete, Nigeria

Corresponding author: Alabi Kayode O., kay4sa@gmail.com

ABSTRACT: *Credit scoring has become the source of many customers intake to the banking industry, due to the large intake of customers and huge database in the organisation credit default risk as become a challenge since the inception of the financial industry. Datamining is a capable aspect of analysing data aimed to remove beneficial information from incredible numbers of multifaceted dataset. In this study an effective prediction technique that helps the financial industry to predict the credit eligibility for customers applying for loan is proposed using the Chi-Square feature selection model and (SVM-RBF) model. Preceding to constructing the model, the data set is pre-processed, reduced and deliver effective predictions. The results shows that chi-square and SVM-RBF model accuracy was 80.29% and compared with the-state-of-art.*

KEYWORDS: *Churn Prediction, Bank Credit, SVM-RBF, Chi-square, Datamining, Credit scoring*

1. INTRODUCTION

Credit scoring has grown into the foundation of many customers intake to the banking industry, due to the large intake of customers and huge database in the organisation credit default risk as become a challenge since the inception of the financial industry. Credit scoring is a method of predicting or assessing whether an applications would be capable of repaying the loan or withdraw from the organisation [6].

Organisations need to regulate the rate at which default increases in the financial industry. The aims of solving this problem is by using dimensionality reduction techniques in removing the irrelevant data, moderate costs and increase the interpretability and performance [5].

Many methods have been proposed by different researchers such as Artificial Neural Networks (ANNs), Multilayer Perceptron (MLP) and k-Nearest-Neighbor (kNN) algorithms so as to improve the decision-making. In recent times, several approaches like stochastic optimization technique, evolutionary algorithms and support vector machine have revealed capable results in terms of accuracy prediction [9].

This study performs Chi-Square feature selection model to fetch relevant information in a huge dimensional data and Support Vector Machine as the classification task to differentiate the possible debtors from the non-debtors and thereby trying to reduce the bad debt [1].

2. GENERAL INFORMATION

Customer Relationship Management (CRM) datamining was suggested, using supervised learning approach, Decision Tree, executed with CART algorithm was proposed for customer retention procedure [8].

A sector applied the data mining is Customer Retention in banking depending on different approaches such as Value estimate approaches and Classification approaches, applications of credits load can be categorize in diverse ways, it efforts to calculate predictable number for novel uses credit Neural Network and regression was implemented, and datamining was proposed for customer descriptive and predictive purpose.

Association rules obtains the association among the database attributes on the importance to develop a multi-attribute correlation, sustaining maintenance and reliance of every threshold. Sequential categorises interactions amid times of objects deliberated as association detection on the sequential database.

[1] proposed and predict the status of loans for banking sector database, using three classification algorithms: bayesNet, j48 and naiveBayes, using Weka. J48 selected the best algorithm accuracy.

A prediction of clustering Algorithm that is Multidimensional was applied by [4] to identify applicants with poor credit loan request. The Stages of loan valuations were introduced to evade idleness, an association Rule was integrated, to predict a better accuracy and lesser computational time.

[3] used Logistic Regression and Support Vector Machine algorithms, including Linear and Non-Linear Deep Neural Networks, were proposed to loan applicant in order to reduce creditor acceptance

of loans and predict the probability of defaulting of issued credits. Two phase model was implemented; the first predictive phase was loan rejection, while the second phase predicts credit risk for loans approved. Logistic Regression performer much better in the first phase, with 77.4% score of test set recall macro. The second phase only uses Deep Neural Networks, were the best performance achieved was validated with 72% set recall score, for creditors. This results stated that AI can boost with present credit risk models by 70% decreasing the default rate of loans issued. The investigation algorithm can be developed in times of Sensitivity, Accuracy and Performance matrix.

3. PROPOSED MODEL

Monthly churn rate for bank customers were used to determine the extent of customer churn in an Australian bank. The customers' perception about causes of churn was gathered through an open source which was analysed with WEKA. To develop the predictive model an already defined churn dataset with 690 records of 15 fields were made available freely by an Australian bank. Data mining systems is proposed to improve the model with WEKA, the chi-square feature selection was used to fetch the relevant information in the data and SVM-RBF classifies were used to derive the propensity and show the characteristics of those who churned for a customer in the dataset to churn.

Data Collection

This is the way information is collected from the sample. The data was gathered through an opened source to find out if customers will churn. This study uses an Australian banking data record, from UCI Warehouse. The dataset contains non-missing values and fits to a high-dimensional dataset with 15 attributes and 619 instances. The dataset comprises of 619 data, 15 attributes with class qualities. The overview of the dataset and the definition of its attributes are shown in Table 3.1, below. The dataset is available in WEKA's .arff format. The credit of customers' dataset was employed and preprocessed using .csv in Microsoft excel to conform to the WEKA tool environment.

Table 3.1:Australian Credit Data Description

Number	Dataset	German Credit Data
1	Attribute type	Category
2	Number of Attribute	15
3	Number of Instance	619
4	Missing Value	No
5	Number of Class	2

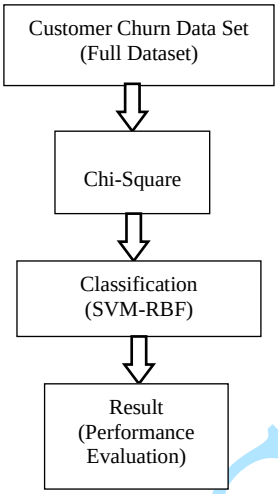

Figure 1. Proposed System Design

4. EQUATIONS ACHIEVEMENT

4.1 Chi-square

An independence chi-square test differentiate two variables in a contingency format which shows if they connected to one another. In general context, it checks that the categorical variables are distributed differently from one another. A very small statistic of the chi square test means that the data observed fit extremely well with the expected data. That is, the relationship exist.

Chi-Square (X^2) is a method of statistics that test two between features of independence variable. Chi-Square is defined as:

$$X^2(t,c) = \sum_{e_i \epsilon 0,1 e_c} \sum_{\epsilon 0,1} \frac{(N_{e_t e_c} - E_{e_t e_c})}{E_{e_t e_c}} \quad (1)$$

Wherever t is a feature in class c, N is the frequency observed, E the frequency expected. e_t is equivalents to 1 if the entity comprises a feature t and e_t equivalents to 0 if the entity does not comprise t. e_c equivalents to 1 if the entity is in class c and e_c equivalents to 0 if the entity is not in class c [Wal16].

4.2 Process of SVM-RFE

SVM classification approach is based on the concept of statistical learning [7]. The feature space of a margin is determined through the hyper-plane of SVM hypotheses, the mapping function Φ was mapped from the input space. Using $\overrightarrow{x_i}$ and $\overrightarrow{z_i}$ to symbolize the feature space and input space are pair of original corresponding vectors, then $\overrightarrow{Z_i} = \Phi(\overrightarrow{X_i})$. A record with a samples d could be denoted as $\{\overrightarrow{X_i}, y_i\}$, $i = 1, 2, \ldots, d$, with $\overrightarrow{X_i}$ $i \in \{0, 1\}^m$, signify a data model, and $y_i \in \{+1, -1\}$, signify a sample of the class label. For a model testing X, the optimum hyper-plane of the feature space constructed by SVM is:

$$< w, \Phi(X) > + b = 0 \quad (2)$$

A problem of optimization needs to fulfil the

restrictions as follows [7]:

$y_i [< w , \Phi(\overrightarrow{X_i}) > +b] + \xi_i - 1 \geq 0 \; \xi_i \geq 0 \; i = 1,2,..., d$ (3)

$$\frac{\min_{w,\xi_i} \|w\|^2}{2} + C\left(\sum_{i=1}^{d} \xi_i\right) \quad (4)$$

It was showed that the hyper-plane fulfils the equation above restrictions of an optimum hyper-plane. However, a fixed constant C governs the interchange between exploiting the margin and reducing the length of the training error. A problem of optimization is normally rendered by the Lagrangian into its double form. This question can be obtained by weight vector and Lagrangian hyper-plane function.

$$w = \sum_{i=1}^{d} a_i y_i < \overrightarrow{Z_I} \quad (5)$$

$$f(Z) = b + \sum_{i=1}^{d} a_i y_i < \overrightarrow{Z_I}, Z > \quad (6)$$

Moreover, <,> means the two vectors is the inner produce. Kernel function $K(\overrightarrow{X_i}, \overrightarrow{X_j})$ computes the two vectors of feature space of an inner product: $K(\overrightarrow{X_i}, \overrightarrow{X_j}) = < \Phi(\overrightarrow{X_i}), \Phi(\overrightarrow{X_j}) > = < \overrightarrow{Z_i}, \overrightarrow{Z_j} >$

If a nonlinear kernel, such as RBF, SIGMOID kernel, is applied to SVM, the weight vector *w* cannot be calculated directly to the equation 4, the mapping function Φ is not known. Linear kernel is mostly adopted in research: $K(\overrightarrow{X_i}, \overrightarrow{X_j}) = < \overrightarrow{X_i}, \overrightarrow{X_j} >$. Weight w for a linear kernel can be defined as:

$$w = \sum_{i=1}^{d} a_i y_i < \overrightarrow{X_I} \quad (7)$$

SVM-RFE using the linear kernel algorithm begins with features and removes a feature using the least squared weight stage by stage till features are all rated. Iterations, w_i^2 are the ranking measure features [7].

The objective function in the SVM-RFE algorithm is $J = \|w\|^2/2$ as used in the OBD algorithm, which estimates the shift of *J* by eliminating the *ith* gene by extending *J* to second order in the Taylor sequence:

$$\Delta J(i) = \frac{\partial J}{\partial wi} \Delta wi + \frac{\partial^2 j}{\partial w_i^2} \Delta w_i^2 .$$

Deleting features having the least squared weight in each iteration causes least effects on *J*. w_i^2 is ranking criterion as adopted and, in order to boost the algorithm 's performance, more eliminated features can be obtain at each step [7].

Evaluation Performance Metrics

The evaluation metrics carried out in terms of classification accuracy, time, specificity, precision and sensitivity, the terms are defined [2].

System Configuration - Hardware Configuration

Processor: Intel Celeron CPU N3060, Speed: 1.60Ghz, Installed memory RAM: 2GB, and Hard Disk: 500 GB.

Configuration Software

Operating System: Windows 10, Programming Tool: WEKA, and Dataset: Australian Bank Analysis of the model.

WEKA is used as a data mining software tool for analysis in carrying out the investigation and achieving a suitable result for credit scoring.

5. EXPERIMENT RESULTS AND PROCEDURES

The data preparation task is done in the user interface environment. The given dataset is pre-processed into the WEKA environment so as to remove noise, inconsistency, bias and redundancies. As shown in figure 5.1 the dataset is loaded in the weka interface, the dataset is selected for classification. The Australian bank dataset consists of 690 attributes and 15 instances with 2 distinct labels consisting on churners and non-churners.

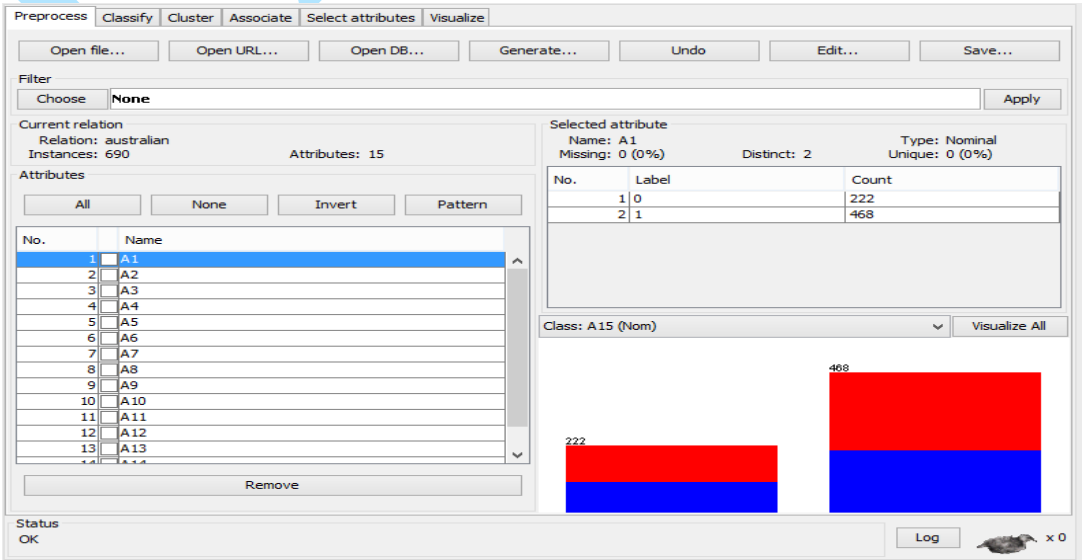

Figure 5.1: Pre-processed Australian Dataset

The dataset is pre-processed and passed into the chi-squared feature selection technique to fetch out relevant information in the dataset which will help to better the output performance. Figure below shows the process involved and in achieving the result in this study.

The selected features were pre-processed into weka environment and it is used for classification using SVM-RBF. The dataset contains 690 instances and 15 attributes.

The selected features are passed to the classifier using SVM-RBF due to its recorded efficient performances in literature. 10-folds cross validation is used to fetch the result of this experiment. Figure 5.2 below shows the process and result output of the performance of this study. The experiment is conducted and the result obtained is shown in the figure below.

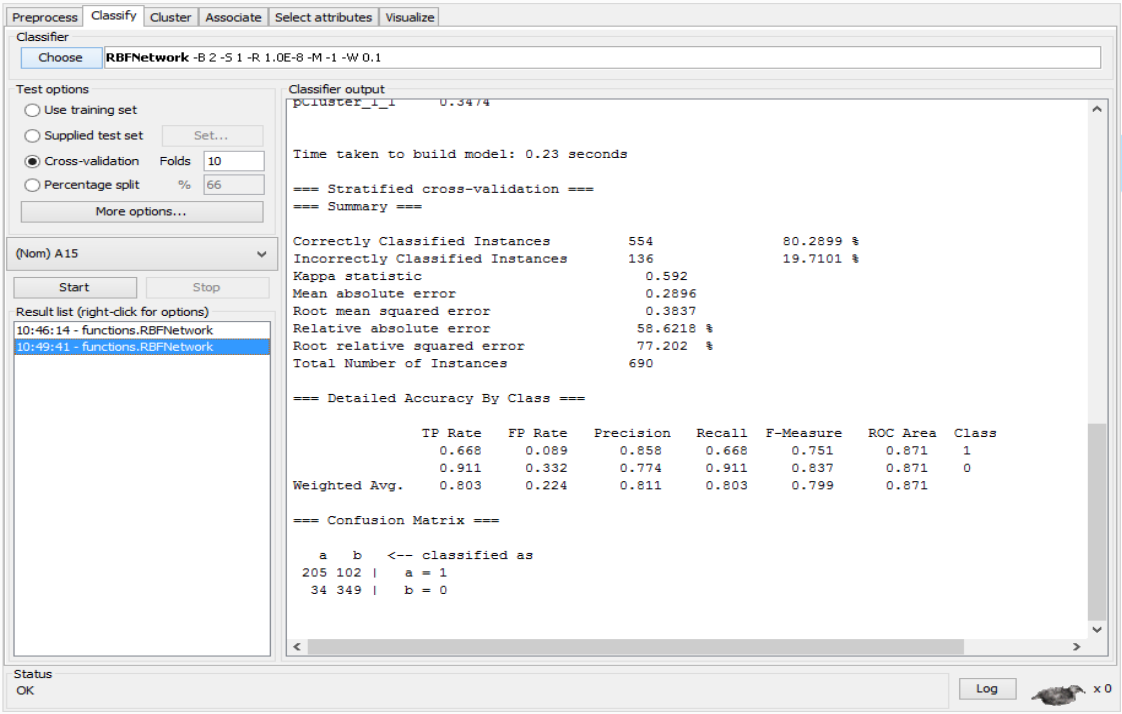

Figure 5.2: Result of the Experiment Chi-Square -SVM-RBF

6. EQUATIONS ACHIEVEMENT

The performance evaluation metrics of Chi-Square and SVM-RBF were calculated as follows:

Correctly Classified instances (accuracy) =

$$\frac{TP+TN}{TP+TN+FP+FN} = \frac{205+349}{205+349+34+102} * 100 = 80.29\%$$

Incorrectly Classified instances =

$$\frac{FP+FN}{N} = \frac{34+102}{690} * 100 = 19.71\%$$

Misclassification rate (mean absolute error) =

$$\frac{FN+FP}{N} = \frac{102+34}{690} = 0.20$$

$$\text{Sensitivity (Recall)} = \frac{TP}{TP+FN} = \frac{205}{205+102} = 0.67$$

$$\text{Specificity} = \frac{TN}{FP+TN} = \frac{349}{34+349} = 0.91$$

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{205}{205+34} = 0.86$$

$$\text{F-measure} = \frac{2*Recall*Precision}{Recall+Precision} = \frac{2*0.67*0.86}{0.67+0.86} = 2$$

The predictive performance was analysed using the confusion matrix and parameters evaluation such as Recall, Precision, F-measure and Accuracy are calculated shown in the below Table 6.1.

Table 6.1 predictive performance summary

Algorithm & Performance	Chi-Square + SVM-RBF
Correctly Classified instances (accuracy)	80.29%
Incorrectly Classified instances	19.71%
Misclassification rate (mean absolute error)	0.20
Sensitivity (Recall)	0.67
Specificity	0.91
Specificity	0.86
F-measure	2

CONCLUSIONS

From this study, feature selection method using chi-square and SVM-RBF classification was used to develop a predictive model in a financial industry using Australian dataset to predict and classify loans

applicant who are eligible or not eligible for loan collection. The model was implemented using Weka application. The result shows that chi-square and SVM-RBF accuracy was 80.29% shown in the predictive performance table. This analysis restricts itself to credit loan estimation, and no measures to contain retention guidelines were examined. The future course of research could be to evaluate the retention guidelines by selecting suitable variables from the dataset.

REFERENCES

[1] **J. H. Aboobyda, M. A. Tarig** - *Developing Prediction Model of Loan Risk in Banks Using Data Mining.* An International Journal (MLAIJ) Vol.3(1):1-9, 2016

[2] **M. O. Arowolo** - *Development of a hybrid dimensionality reduction for microarray dataset classification model*, Vol. 9 (11), pp. 57-63, 2017

[3] **J. D. Turiel, T. Aste** - *P2p loan acceptance and default prediction with Artificial intelligence.* Vol. 1(1): 3-9 2019

[4] **K. Kavitha** - *Clustering Loan Applicants based on Risk Percentage using K-Means Clustering Techniques*, vol 6(2):162-166 2016 – IF ONE AUTHOR

[5] **K. Nikita, L. Stefan, B. Bart** – *Profit-Oriented Feature Selection in Credit Scoring*, Applications DOI: 10.1007/978-3-030-18500-89, 2019

[6] **T. Nupura, S. Pravin** - *Game Predictive Analysis of Credit Score for Credit Card Defaulters*, Volume-7 Issue-5S2 9-12, 2019

[7] **L. Quanzhong, Z. Yang** - *Feature selection for support vector machines with RBF kernel*, DOI: 10.1007/s10462-011-9205-2, 2011

[8] **P. Raju, V. Bai, G. Chaitanya** - *Data mining: Techniques for Enhancing Customer Relationship Management in Banking and Retail Industries*, Vol2(1), 2650- 2657, 2014.

[9] **H. Van-Sang, N. Ha-Nam** - *Credit scoring with a feature selection approach based deep learning*, DOI: 10.1051/matecconf/2016 054 05004 2016

[10] **G. Walaa** - *SVM-Kmeans: Support Vector Machine based on Kmeans Clustering for Breast Cancer Diagnosis*, Vol(5)2: 2-6, Ain Shams University Cairo 2016