Scientific Journal Impact Factor (SJIF) : 3.518
International Journal of Modern Trends in Engineering and Research
ISSN (ONLINE) : 2349 - 9745 ISSN (PRINT) : 2393 - 8161

# A New Method Based on Clustering and Feature Selection for Credit Scoring of Banking Customers

**Seyedeh Maryam Anaei[1] and Mohsen Moradi[2]**

[1]*Department of Computer engineering,Islamic Azad University Boushehr Iran*

[2]*Department of Computer engineering,Islamic Azad University Beyza Iran*

**Abstract—**One of the issues which must be taken into account by credit policy makers in banking industry is risk management. Credit risk is one of the most serious risks faced by banks. This risk may bring about negative consequences such as customers' inability or unwillingness to fulfill their obligations to the bank. To manage and control credit risk, the use customer credit classification system is inevitable. Such system puts customers in appropriate classes based on available data and records .Credit scoring generally aims to provide an accurate prediction of customer competency. To this end, multiple statistical techniques plus artificial intelligence have been used. However, most models are not able to provide multi-class categorizations using two-class data. Nevertheless, a customer is assessed based on some degrees of goodness or badness. Therefore, after removing data noises through clustering, it is usually attempted to use a multi-class support vector machine (SVM) to classify new data. Feature selection and parameter adjustment is performed via genetic algorithm. Compared with other scoring models, the proposed model has been able to improve classification accuracy for German and Australian credit datasets

**Keywords—** credit scoring, clustering, support vector machine (SVM), genetic algorithm

## I. INTRODUCTION

One of the key success factors in lender organizations in general and banks in particular is to assess the borrower's credit position. Credit risk appears if a wrong decision is made when confirming the borrower's request. Therefore, credit risk is one of the challenges that may be faced by financial institutions. Historically, credit risk was assessed through judgmental and subjective decisions made by creditors through analyzing available data. Therefore, wrong decisions concerning to grant or not to grant loans were taken based on a mainly subjective and time-consuming process [1]. As such, credit scoring is a very simple problem of data mining classification. Credit scoring is generally a term used to describe formal statistical methods to classify credit applicants into good and bad groups. In fact, credit scoring aims to divide applicants into two groups: applicants with good credit and those with bad credit [2]. In most studies, a number of models are used which instead of providing two-class predictions, divide customers into several classes using specific criteria or suitable cut-off values. Logit and probit models [3] or neural networks [4] are examples of such methods. However, the selection of suitable cut-off values that can work well under trial and test conditions is a very complicated process. Besides, the use of such values with these models, each with its own weaknesses, is problematic. Logit and probit models, for instance, assume that data are normally distributed and the model used will fit out a normal distribution on data. When it comes to neural networks, the adjustment of initial parameter is an important issue that must be determined in advance. Imbalanced credit data can also be regarded as another serious problem in this regard. In the real world, the number of customers who make a default when repaying their obligations is much less than those who repay their debts on time. To eliminate the above problems, the present study seeks to propose a model which is mainly based on clustering mechanism. To improve the model performance and precision, the model parameter and features are selected simultaneously by genetic algorithm to provide a good model for credit scoring of banking customer.

 

## II. LITERATURE REVIEW

### A. Credit Scoring

Credit scoring is a study to assess the possibility of granting a loan. The aim of credit scoring is to find out if the borrower is able to do useful economic activities in the market or not and whether he is able to repay the loan on due. Credit scoring is usually done by the bank accountant who is possibly a member of the evaluation committee. Such assessment is performed to analyze all factors involving credit granting such as business credit efficiency and financial rating of borrowers. Six types of analysis are performed for the purpose of credit scoring [5]:

1. **Character:** Character refers to the borrower's credit rating which measures the possibility that the borrower repay his debts and his ability to do so bed on a defined agreement. Character is the most important factor in credit scoring.
2. **Capital:** It refers to initial assets possessed by the borrower. The greater the assets, the more likely will be granting a loan.
3. **Competence:** Competence is the borrower's ability to pay installments. It measures income and expected income level during the repayment period. Competence is needed for the assessment of the borrower's repayment ability.
4. **Collateral:** It refers to the borrower's property which must be given to the bank as a loan guarantee. The value of collateral must be the same as the value of the loan granted to the borrower.
5. **Economic conditions:** They refer to the society's potentials, culture, and the borrower's economic-business status.
6. **Constraints:** They include all restrictions and barriers that prevent the continuity of economic activities such as regulations and limited resources [5].

### B. Support Vector Machine

Support vector machines (SVMs) are one of the best algorithms to buy, sell, or classify small samples and used for regression analysis. When SVMs applied to improve classification or used in regression analysis, three main problems may occur: 1) Selection of optimal features, 2) Core selection, and 3) Determining core parameters. In order to improve classification precision, parameters must be optimized [6].

### C. Clustering

Clustering is one of the most important processes in data analysis and pattern discovery which contribute to locating natural boundaries in data. Clustering can be introduced in two ways: 1) A dissimilar n*n matrix is given, 2) An n*d matrix is provided and its both rows are described by on object. Algorithm outputs may be presented in two forms: 1) Grouping objects into discrete sets, 2) Hierarchical clustering where a tree is used to classify objects. The first type algorithms are quicker to be used with a time of O(nd) compared with time of $O(n^2 \log(n))$ required for hierarchical clustering. One of the well-known algorithms used for clustering is K-means.

### D. Feature selection

Feature selection –called also feature reduction or variable selection- is a machine learning technique. Feature selection will improve the efficiency of the learning model by removing redundant or irrelevant features in the dataset. Feature selection is normally divided into two main groups: Feature rating and selecting a subset of features.

In feature rating where features are rated based on a specific criterion, the features with the minimum scores will be removed. In contrast, an optimal set of features are searched and identified by feature subset selection. This may be done in three ways: Filtering, packaging, and insertion [10].

 

## III. DATASET

The data used in this study were collected from German and Australian credit data which included real customers. The German dataset consisted of 1000 records with 24 features and the Australian dataset included 690 records with 14 nominal and numerical features. Table 1 shows the number of samples available in the Germany and Australian data that are classified into good and bad customers.

TABLE 1: Description of data used in the study

| Credit Dataset | Number of Samples | Number of Features | Good samples | Bad samples |
|---|---|---|---|---|
| **Germany** | 1000 | 24 | 700 | 300 |
| **Australia** | 690 | 14 | 307 | 383 |

As was mentioned, the German credit data included 24 features as follows: Credit position and how installments are repaid, the applicant's age, gender, number of children, education level, the partners' education level, job experience, monthly income, current assets and properties, current housing, duration of residing in the current housing, the amount of loans received, collateral value, duration of installment payment, the goal of receiving loans, the history of dishonored cheques, credit history of granting banks or institutions, duration of having relationship with the bank, saving account balance, current account balance, loan number, and loan date. The features in the Australian dataset included: gender, age, marital status, job, current address, housing, home phone number, duration per month, loan date, loan amount, job experience (in years), number of loan received, and loan number.

## IV. PROPOSED ALGORITHM

Fig. 1 shows different steps of implementing the proposed model. The proposed method consists of two main steps: Model preparation and model construction. In the first step, the problems discussed in the introduction part of the study are taken into account and appropriate solutions to them are suggested. The next sections deal with steps taken to construct the model.

### *A. Data preparation*

Appropriate solutions to the above problems are introduced in this section:
First, all data in the both sets are normalized at interval (0, 1) using Eq. (1) in order to improve classification accuracy:

$$Normalized\ (xi) = \frac{x_i - Xmin}{Xmax\ |-Xmin} \qquad (1)$$

Where, xi is the datum that is to be normalized, Xmin is the smallest datum, and Xmax is the greatest datum. Normalization is performed on numerical data while nominal data remain unchanged.
In the next stage, the data noise is removed by K-means clustering. To do so, the data mean for each cluster is calculated. If the mean is greater than the total mean, the data are removed, i.e. they will be placed at another cluster. The data noises in the stage are removed for learning purposes but they are included in the final test. K-means algorithm used in the model is shown as follows. The number of clusters (c) has been determined in advance. The target function is written as follows (Eq. 2):

$$J = \sum_{j=1}^{c}\sum_{k=1}^{n} u_{ik}^{m} d_{ik}^{2} = \sum_{j=1}^{c}\sum_{k=1}^{n} u_{ik}^{m} \left\| x_k - v_i \right\|^2 \qquad (2)$$

Where, m is a real number greater than 1. In most cases, m equals 2. In addition, Xk is sample K*th*, Vi is the center of cluster i, Uik is the belonging level of sample i*th* in cluster k, and ‖*‖ is the similarity of the sample form the center of cluster.

*B. Model construction*

According to results from various studies, if the two steps of feature selection and regulation of SVM parameters are performed simultaneously, the performance will be improved greatly compared with cases where the two steps are performed separately. Therefore, the optimal values of SVM parameters and important features can be determined simultaneously by using genetic algorithm.
Model construction algorithm includes the following steps:

1. Data normalization
2. Classification by using K-means algorithm
3. Determining average intervals in each cluster to detect data noises
4. Noise elimination
5. Feature reduction using PCA
6. Feature selection using GA
7. SVM classification through RBF core function

The core of the proposed model is a multi-class SVM classifier that is supposed to classify test sample with a high precision after training. Here, a RBF core function with a 14.51 mm radius is used. The genetic algorithm used in this study is a binary algorithm. It does not consider features with a value of zero while it selects features with a value of 1 as optimal features. The efficiency of the selected features will be evaluated by a neural network. Given the impact of features on data classification, a score is assigned to each feature. Finally, the assigned scores are used to make decisions. The number of chromosomes for each generation is 100, mutation rate is 0.02, and the number of generation is 50 as determined by one-point and two-point remix method. According to this method, parents are produced in three ways: randomly, roulette wheel, and racing. To eliminate the problem of two-class nature of the data using clustering data, the data are labeled in similar groups in K-means clustering stage.

SVM usually classifies the data into two groups. However, in this algorithm; multi SVM is used to classify the data into four groups; groups 1 and 2 include bad and relatively bad customers while groups 3 and 4 consist of good and relatively good customers. As such, loans are granted based on a score assigned to each customer. Therefore, a customer will not be placed mistakenly in bad or good groups.

The main steps in SVM are summarized as follows:

1. Preparing a data matrix
2. Selecting an appropriate core function
3. Selecting core function parameters
4. Implementing training algorithm using RBF function as follows (Eq. 3):

$$K(x, y) = \exp^{\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)} \quad (3)$$

5. Classification of test data by SVM

C. RESULTD

The model accuracy can be determined using test data in the test stage in terms of its ability to identify the four groups. K-fold cross validation with k = 10 was used to assess the algorithm. The criteria used to compare the results are defined as follows:

1. Accuracy = (TP+TN)/(P+N)
2. Precision = TP/(TP+FP)
3. Recall = TP/(TP+FN)

4. F.measure=(2*Precision*Recall)/(Precision+Recall)

TABLE 2: Assessment of the proposed model

| Criterion | Germany | Australia |
|---|---|---|
| **Accuracy** | 90% | 94.2% |
| **Precision** | 0.95% | 0.94% |
| **Recall** | 0.95% | 0.98% |
| **F-measure** | 0.95% | 0.95% |

Given that several techniques have been used in this study, comparisons are made in the two following modes:

1. Similar classification without preprocessing and feature selection:
   The proposed algorithm is compared with LibSVM algorithm and the hybrid SVM+GA algorithm is compared with SVM algorithm as shown in Table 3:

TABLE 3: Comparing algorithms

| Algorithm | Accuracy | |
|---|---|---|
| | Germany | Australia |
| **LibSVM** | 79.1% | 87% |
| **SVM+GA** | 77.92% | 86.9% |
| **SVM** | 70.5% | 81.58% |
| **Proposed algorithm** | 90% | 94.2% |

2. Different classification with preprocessing and feature selection
   The proposed model is compared with hybriddecision tree scoring model based on genetic algorithm and K-means algorithm as shown in Table 4:

TABLE4. Comparing algorithms

| Algorithm | Accuracy | |
|---|---|---|
| | Germany | Australia |
| **Hybrid decision-making tree** | 77.76% | 89.33% |
| **Proposed model** | 90% | 94.2% |

## V. CONCLUSION

According to the findings of this study, it can be suggested that the accuracy of the proposed algorithm in terms of data classification is greater than other models and it reduces classification errors. Therefore, the proposed algorithm is more efficient than other methods and it can be used to assess the credit position of banking customers. In addition, genetic algorithm has a better performance in terms of precision and data distribution. However, the great number of calculations used in this method should not be disregarded as it may be a take a lot of time. Of course, this is acceptable because of off-line nature of credit scoring. The proposed algorithm can be used as an alternative to SVM for future applications. Besides, every classifier model with multi-class grouping ability can be employed to for data classification purposes. Other models of revolutionary algorithms such as Bird Flocks and Ant Colony can be used to optimize data classification. Finally, simple fuzzy logic can be used as it improves both the ultimate accuracy and interpretability of the final model.

## REFERENCES

[1] Baesens, B., Gestel, B., Viane, T., Stepanova, S., Suykens, M., &Vanthienen, J. (2003).Benchmarking State-of the-artClassification Algorithms for Credit Scoring. Journal of the Operational Research Society , PP.627- 635,2003.
[2] Chen, W., Xiang, G., Liu, Y., &i Wang, K. Credit riskEvaluation by hybrid data mining technique. SystemsEngineering Procedia3 , PP.194- 200,2012

[3] Kambal, E., Osman, I., Taha, M., Mohammed, N., &Mohammed, S. Credit Scoring Using Data Mining Techniqueswith Particular Reference to Sudanese International Conference on Computer, Electrical and Electronics Engineering (ICCEEE) ,PP.378-383,2013.
[4] Westgaard, S., &Wijst, N. Default Probabilities in a Corporate Bank Portfolio: A Logistic Model Approach.European Journal of Operational Research , 338,2001.
[5] I GustiNgurahNarindraMandalaa, Catharina BadraNawangpalupia*, FransiscusRianPraktiktoa / InternationalJournal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 2, Issue 2, pp.738-742, Mar-Apr 2012.
[6] Weimin Chen, Guocheng Xiang, Youjin Liu, Kexi Wang. Credit risk Evaluation by hybrid data mining technique, SystemsEngineering Procedia 3 PP.194 – 200, 2012.
[7] Cheng-Lung Huang, Mu-Chen Chen, Chieh-Jen Wang,"Credit scoring with a data mining approach based on supportvector machines", Expert Systems withApplications,33(4),pp.847-856,2007.
[8] Defu Zhang, Stephen C.H. Leung, Zhimei Ye. A Decision Tree Scoring Model Based on Genetic Algorithm and KmeansAlgorithm, Third 2008 International Conference on Convergence and Hybrid Information Technology.
[9] Shin-Chen Huang, Min-Yuh Day. A Comparative Study of Data Mining Techniques for Credit Scoring in Banking, IEEE IRI 2013,Agust 14-16, San Francisco,California, USA978-1-4799-10502/13/$31.00 © 2013 IEEE.
[10] L. A. Blum, P. Lngley, "Selection of Relevant Features andExamples in Machine Learning". Artificial Intell, 1997, pp. 245-271.
[11] L1 Graph based on sparse coding for Feature Selection, Lecture Notes in Computer Science, vol 7951, pp. 594-601,2013.
[12] Jayasree V, Vijayalakshmi R.(2013). A Review on Data Mining in Banking Sector, American Journal of Applied Sciences, 1160-1165.
[13] Basel Committee on Banking Supervision: International Convergence of Capital Measurement and Capital Standards: ARevised Framework. (2005).Bank for International Settlements
[14] Weimin Chen, Guocheng Xiang, Youjin Liu, Kexi Wang. (2012 ). Credit risk Evaluation by hybrid data mining technique,Systems Engineering Procedia 3,194 – 200.
[15] Kamalloo,E., SanieeAbadeh,M.(2010). An Artificial Immune System for Extracting Fuzzy Rules in CreditScoring,IEE.