ISSN No. 0976-5697

Volume 4, No. 9, July-August 2013

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

A Novel Method for Classification of Bank Customers Based on the Rough Set and Rules Extraction

Ali Reza Taleghani
Department of Electrical and Computer &it Engineering
Qazvin Branch, Islamic Azad University
Qazvin, Iran
taleghani.a.r@gmail.com

Mohammad Saniee Abadeh
Faculty of Electrical and Computer Engineering
Tarbiat Modares University
Tehran, Iran
saniee@modares.ac.ir

Abstract: In recent years, credit scoring studies have been considered by researchers. In this paper, with a new approach based on statistical methods and rough set theory, the rules are extracted for classification. To this aim, the credit scoring data set of UCI University, Australia has been used. The proposed algorithm is very simple and comprehensible and is very efficient for the problem under study, in comparison to other known methods.

Keywords: credit scoring; classification; rough set; statistics

I. INTRODUCTION

It is a very important issue to judge the credit condition of consumers in the credit industry. With the rapid growth in this field, credit scoring models have been widely used for the credit admission evaluation. The credit scoring models are developed to distinguish which customers are belong to good or bad class with their related attributes such as income, marital status,age or based on the past records. Most credit scoring models have been widely developed by reducing redundant features to improve the accuracy of credit scoring models during the past few years [1].

The purpose of credit scoring model is to classify credit applicants to either a good credit group that is probable to pay back financial obligation or a bad credit group who has high risk of defaulting or becoming delinquent on the financial obligation [2]. It is one of the earliest financial risk management tools developed [3]. Its significance is more highlighted because of recent financial crisis.

II. PREVIOUS WORKS

Accuracy and transparency are two important criteria that should be satisfied by any credit scoring system. Good accuracy enables correct assessment and thus avoiding any heavy losses associated with wrong predictions while transparency enables financial analysis to understand the decision process. Statistical methods such as Linear discriminant analysis (LDA) and logistic regression (LR) are the most commonly used methods in building credit scoring models. However, artificial intelligence like neural networks and genetic algorithms provide a new alternative to statistical methods in building non-linear, complex and real world systems. Furthermore, techniques using neural networks and genetic algorithms have reported to have achieved higher prediction accuracy than those using LDA and logistic regression and others methods [4],[5], [6], [7], [8].

The study is organized in the following way; in section 2, the research methodology is presented. In sections 3 and 4, the two proposed algorithm are described in more detail. Section 5 presents the results of simulation, and the final section is devoted to conclusion.

III. CREDIT SCORING DECISION PROBLEM

Before credit scoring models came into wide use in 1980, human judgment was the sole factor in making decisions who are the good and bad applicants, and then who receive credit. Judgmental method was not only slow but also unreliable because of the human error and bias.

Credit scoring models nowadays are based on statistical or operation research methods. These models are built using payment historical information from thousands of actual consumers. Credit scoring objective is to assign credit applicants to either good customers or bad customers. Therefore credit scoring lies in the domain of the classification problem [9]. The credit scoring model captures the relationship between the historical information and future credit performance. This relation can be described mathematically as follows:

$$f(x_1, x_2,...,x_m) = y_n \qquad (1)$$

Where each customer contains attributes: :x_1, x_2,..., x_m, y_j , denotes the type of customer, for example good or bad. is the function or the credit scoring model that maps between the customer features (inputs) and his creditworthiness (output). The task of the credit scoring model (function) is to predict the value of i.e. the creditworthy of customer i by knowing the .i.e. the customer features such as: income, age. Many methods have been suggested to develop credit scoring models but the most popular methods adopted in the credit scoring industry are linear discriminant and logistic regression and their variations [8].

A. Rough sets theory approach

Rough sets theory (RST) is a mathematical tool that had been used successfully to discover data dependencies and reduce the number of attributes contained in a data set by purely structural methods. RST was firstproposed by Pawlak [10] to deal with vagueness or uncertainty. Rough sets do not

need any pre-assumptions or preliminary information about the data. One attribute is chosen as the decision variable and the rest of them are the condition attributes. Two partitions are formed in the mining process. The approach is based on the refusing certain set boundaries, implying that every set will be defined using a lower and an upper approximation. As can be observed from Fig. 1, the object that belongs to a set with certainty is called lower approximation while upper approximation contains all objects that may possibly belong to the set. Decision rules derived from lower approximation represents certain rules as well as extracted from upper approximation corresponds to possible rules. An important issue in the RST is about feature reduction based on reduct concept. A reduct is a minimal set of attributes B # A such that IND(B) = IND(A), where IND(X) is called the X-indiscernibility relation. In other words, a reduct is a minimal set of attributes from A that preserves the partitioning of universe and hence the ability to perform classi fications. RST has been successfully applied to real-world classification problems in a variety of areas, such as pattern recognition. Wang and his colleagues proposed a new feature selection strategy based on rough sets and particle swarm optimization [11]. Zhao and his colleagues also made an empirical experiment for letter recognition for demonstrating the usefulness of the discussed relations and reducts [12]. There are many other rough sets algorithms for feature selection. The basic solution to finding minimal reducts is to generate all possible reducts and choose any with minimal cardinality, which can be done by constructing a kind of discernibility function from the dataset and simplifying it. However, this is time consuming and therefore is only practical for simple datasets. Finding minimal reducts or all reducts has been shown as NP-hard problems [13].

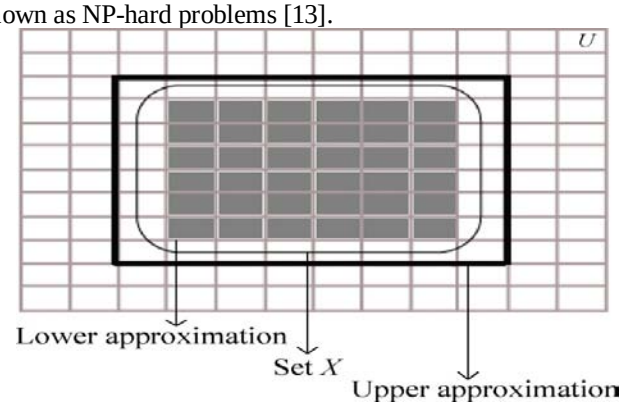

Figure 1. Rough sets approximation

IV. METHODOLOGY

The first proposed algorithm, using rough set theory (low approximation), considers the features which highly distinguish classes in the population, and selects the features that can classify the train data.

In the second proposed algorithm, using the approach of the first algorithm, and considering the level of discrepancy in each feature, selects the value of the feature or marks it as don't care. In fact, the first algorithms considers the distinction among features, while the second algorithm, in addition, pays attention to each feature independently and considers it for classification. The main point about these two algorithms is the method of discretizing numerical values to intervals to obtain the most distinction among classes. To this aim, first, the effective features are selected through classifier logistic + genetic method. In this way, 7 effective features were selected, and then were discretized in a supervised method. The tool utilised was weka.

Also, by reducing the probability of selecting train data that were covered by the obtained rules, a better search is conducted.

A. *The problem of classifying pattern by rules*

Classifying patterns is a problem with n dimensions, c classes, and m train patterns. p=1,2,...,m, $X_p = (x_{p1}, x_{p2}, ...x_{pn})$; $A_{j1},...,A_{jn}$ is the nominal values of features which are in the form of discretized intervals. To identify the test sample class, the number of rules of each class that has the highest compatibility with the sample is considered as its class.

$$\text{Rule } R_j\text{: If } x_1 \text{ is } A_{j1} \text{ and } \ldots \text{and } x_n \text{ is } A_{jn} \text{ then Class } C_j \quad (2)$$

B. *Coefficient of variation*

In probability theory and statistics, the coefficient of variation (CV) is a normalized measure of dispersion of a probability distribution. It is also known as unitized risk or the variation coefficient. The absolute value of the CV is sometimes known as relative standard deviation (RSD), which is expressed as a CV should not be used interchangeably with RSD (i.e. one term should be used consistently).

The coefficient of variation (CV) is defined as the ratio of the standard deviation δ to the mean μ:

$$CV = \delta/\mu. \quad (3)$$

In selecting or not selecting the feature values, a Coefficient of variation was offered. This criterion, considering the concept of variance showing the discrepancy of discretized interval of every feature is obtained from (3). α is a parameter to be set and show the value of don't care. In the following, the two proposed algorithms are explained according to discrepancy criterion and rough set.

C. *The first proposed algorithm*

In this algorithm, using rough set theory, Table I shows the parameters specifications of first proposed algorithm for data set in our experiments. the following steps were taken (Fig.2):

1. Pre-processing: first, the data set were discretized in a supervised method with the help of weka software.
2. The train set is separated for each class and learning is done separately for each class. The following stages are repeated until N-rule is created.
3. The primary population for each class: k sample is selected randomly. k is also a random number (k<<n, and n is the number of train samples).
4. If the maximum of feature value in population of class0 is the opposite to the feature value in population of class1, then, this value is added to the rule, otherwise, it is marked don't care.
5. Evaluation: the rule is evaluated; if it covers more than Minfitness (minimum fitness) of the train data, it is selected; otherwise, the steps of rule generation are repeated. After running the

algorithm and extracting N/2 rule for each class, N rule is obtained.

6. The probability of selecting covered samples by the generated rules is lowered and the exploratory power of the algorithm is increased.

Evaluating the rules

Fitness of each rule is equal to the number of train samples covered by that rule minus the number of samples wrongly covered. This is calculated by (4). The rate of classifying number of samples correctly classified is obtained from (5).

Fitness(R_i)= number of correctly classify - number of incorrectly classify (4)

Classification rate=(TP+TN)/(TP+TN+FN+FP) (5)

TP = Number of examples satisfying A and C.

FP = Number of examples satisfying A but not C.

FN=Number of examples not satisfying A but satisfying

TN = Number of examples not satisfying A not C.

While (N-rule>0)
Begin
1- Generate initial population For Each Class.
2-IF maximum Frequency of Feature Values in populationclass0! = maximum Frequency of Feature Values in populationclass1 Then add to Rule
Else Don't Care
3.Rule Learning: IF Fitness($Rule_i$)>Minfitness Then Add To Rule Set and N-rule= N-rule-1;
4. Reduce selection probability Samples Of Data Set That Covered With The Rule
End.

Figure 2. An overview first proposed algorithm.

D. The second proposed algorithm

This algorithm is based on the level of Coefficient of variation of feature values. while it is independently calculated for every feature of discrepancy, it is also considered indirectly due to selection in the limited population (Fig.2). If variance is decreased, the probability of selecting the value of the feature with the most repetition in the population is increased. In this method, it is worked according to the first algorithm; besides, the equation (1) of rules is

While (N-rule>0)
BEGIN
1- Generate initial population For Each Class.
2-IF maximum Frequency of Feature Values in $population_{class0}$! = maximum Frequency of Feature Values in $population_{class1}$ Then
IF Variance (Feature) <α*Mean (Feature) Then Don't Care Else add to Rule
Else Don't Care
4-Rule Learning: IF Fitness ($Rule_i$)> Minfitness Then Add to Rule Set and N-rule= N-rule-1;
5-Reduce selection probability Samples Of Data Set That Covered With The New Rule
END

Figure 3. An overview second proposed algorithm.

selected based on Coefficient of variation. Indeed, considering the difference among the values in each class means that this feature has more distinguishing power relative to others. The parameters of this algorithm are presented in Table II.

V. SOME COMMON MISTAKES EXPERIMENTAL RESULTS

Credit data sets in the real-world include various attributes. Two real world data sets were selected for this research, i.e. the Australian and German credit data sets derived from the UCI Repository of Machine Learning Databases. The Australian data set consists of 307 ''good" applicants and 383 ''bad" ones. For each applicant contains 15 features, including 6 nominal, 8 numeric attributes and the final one is class label (good or bad credit). These attributes names have been changed to meaningless symbolic data for the confidential reason.

we evaluate the accuracy of algorithms with 10-fold cross validation (CV) technique. The results in Table III indicate that the proposed method is more efficient in such problems.

Table I. PARAMETER SPECIFICATIONS OF first algorithm

Parameter	*Value*
population size	*3<k<55*
(N-rule) RuleSetSize	*10*
Minfitness	*33*

Table II. PARAMETER SPECIFICATIONS OF second algorithm

Parameter	*Value*
population size	*3<k<55*
(N-rule) RuleSetSize	*10*
Minfitness	*33*

Table III. COMPARISON OF CORRECT PREDICTION ACCURACY OF CCSFAISWITH OTHER CLASSIFIERS FOR AUSTRALIAN DATASET

Algorithm	*Classification Rate*
Discriminant analysis [18]	71.4%
Logistic regression [18]	73.4%
Back propagation neural networks [18]	73.7%
Hybrid neural discriminant model [18]	77.0%
Quadsic [17]	79.3%
CN2 [17]	79.6%
ALLOC80 [17]	79.9%
LVQ [17]	80.3%
CCS-FAIS[14]	80.7%
CCS-FAIS[14]	82.7±0.1%
our Proposed algorithm1	85.5±3.5%
our Proposed algorithm2	

VI. CONCLUSION

This method has the advantage of rule of extraction for an expert system. It has the advantage of being model

comprehensible in comparison to methods such as neural network and SVM. The low number of adjustable parameters in this method is simpler in comparison to the high number of parameters of evolutionary algorithm. In future studies, the use of fuzzy logic and evolutionary strategy for improving its comprehensibility and efficiency is discussed. Also, it is addressed for problems with more than two classes.

VII. REFERENCES

[1] Li. Feng-Chia, "The Hybrid Credit Scoring Model based on KNN Classifier," Sixth International Conference on Fuzzy Systems and Knowledge Discovery, 2009.

[2] Defu Zhang, M. Hifi, Qingshan Chen, and Weiguo Ye, "A Hybrid Credit Scoring Model Based on Genetic Programming and Support Vector Machines," Natural Computation, 2008. ICNC '08. Fourth International Conference on, pp. 8-12 , 2008.

[3] L.C. Thomas, D.B. Edelman, and J.N. Crook, Credit scoring and its applications, Society for Industrial Mathematics, 2002.

[4] D. West, "Neural network credit scoring models," Computers and Operations Research, vol 27, pp. 1131–1152, 2000.

[5] V.S. Desai, J.N. Crook and G.A. Overstreet, "A comparison of neural networks and linear scoring models in the credit union environment," European Journal of Operational Research, 95, pp. 24–37 ,1996.

[6] H.L. Jensen , "Using neural networks for credit scoring," Managerial Finance, vol 18, pp. 15–26 ,1992.

[7] S. Piramuthu ,"Financial credit-risk evaluation with neural and neurofuzzy systems," European Journal of Operational Research, 112, pp. 310–321 ,1999.

[8] V. Desai, J. Crook, and G. Overstreet, "Credit scoring models in the credit union environment using neural networks and genetic algorithms," IMA Journal of Mathematics Applied in Business and Industry, 8(4), pp. 324–346 ,1997.

[9] S. Piramuthu, "Financial credit-risk evaluation with neural and neurofuzzy systems," European Journal of Operational Research, 112, pp. 310–321 ,1999.

[10] Pawlak, Rough classification, Academic Press Ltd ,Vol. 20, pp. 469–483, 1984.

[11] X. Wang, J. Yang, X. Teng, W. Xia, and R. Jensen, "Feature selection based on rough sets and particle swarm optimization," Pattern Recognition Letters, 28 (4), pp. 459–471,2007.

[12] Y. Zhao, Y. Yao, and F. Luo, "Data analysis based on discernibility and indiscernibility," Information Sciences, 177(22), pp. 4959–4976,2007.

[13] A.Skowron, and C. Rauszer, The discernibility matrices and functions in information systems. In R. Slowinski (Ed.), Intelligent decision support Handbook of applications and advances of the rough sets theory Kluwer:Academic Publishers, pp. 311–362, 1992.

[14] Ehsan Kamalloo, Mohammad Saniee Abadeh, "Comprehensible Credit Scoring with Fuzzy Artificial Immune System," ICEE 2010, May 2010.

[15] Lei Shi, Lei Xi, Xinming Ma, Mei Weng, Xiaohong Hu, "A novel ensemble algorithm for biomedical classification based on Ant Colony Optimization," March 2011.

[16] A. Asuncion and D. Newman, "UCI Machine Learning Repository [http://archive.ics.uci.edu/ml/ml.html]," 2007.

[17] K. Leung and F. Cheong, "Consumer credit scoring using an artificial immune system algorithm," *Evolutionary Computation, 2007. CEC 2007. IEEE Congress on*, 2007, pp. 3377-3384.

[18] T. Lee, C. Chiu, C. Lu, and I. Chen, "Credit scoring using the hybrid neural discriminant technique," *Expert Systems with Applications*,vol. 23, Oct. 2002, pp. 245-254.