ISSN: 2321-7782 (Online)
Volume 4, Issue 2, February 2016
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Classifying Data and Predicting Risk towards Multi - Dimensional Dataset using K-Means Clustering Algorithm

Dr. K. Kavitha
Assistant Professor, Department of Computer Science
Mother Teresa Women's University, Kodaikanal - India

Abstract: ***Data classification and prediction are the key techniques in data mining. These concepts are used to classify the data based on the criteria's and group the similar items from large voluminous datasets. Risk assessment is a critical task in banking sector towards identifying the credit risk based on the customer's status. Many researchers have proposed algorithm for assessing the risks in an improved manner but still it has some limitations for evaluation. An improvised risk evaluation of Multi-dimensional Risk prediction clustering Algorithm is proposed .The proposed method overcomes the limitations and integrated K-means clustering algorithm for grouping the good and bad customers separately. Association Rule algorithm is integrated to predict the rules effectively.***

Keywords: Risk Prediction, Clustering, Redundancy, Data Mining, Feature Extraction.

I. INTRODUCTION

The key idea of data mining techniques is to classify the customer data according o the posterior probability. Here it is used to perform the classification and prediction of loan. With the continuous development and changing in the credit industry, credit products play an important role in the economy. Credit risk evaluation decisions are crucial for financial institutions due to high risks associated with inappropriate credit decisions that may result in major losses. It is an even more important task today as financial institutions have been experiencing serious challenges and competition during the past decade. It concerns those lenders to limit potential default risks, screening the customer's financial history and financial background. Banks should control credit management thoroughly. Sanctioning of loan needs the use of huge data and substantial processing time. Before granting loans, banks have to take various precautions such as performance of the firm by analyzing last year's financial statements and history of the customer. The decisions of sanctioning loans may become wrong and resulted in credit defaults. An intelligent information system that is based on clustering algorithm will provide managers with added information, to reduce the uncertainty of the decision outcome to enhance banking service quality.

Due to high competition in the business field, customer relationship management has to be considered in the enterprise. Here analyze the massive volume of data and classify on the customer behaviours and prediction. Customer relationship management is mainly used in banking areas. Data mining provides many technologies to analyze mass volume of data and detect hidden patterns to convert raw data into valuable information. It is a powerful new technology with great potential to help banks focus on the most information in their data warehouse.

Rest of this paper is structured as below: In section 2, research works related to the risk assessment in banks are discussed. The detailed explanations of the proposed framework are given in section 3. Experimental results are reported in the section 4 to prove the efficiency and accuracy of the proposed framework. Finally, section 5 concludes this paper along with directions for future work.

II. Related Work

Credit risk evaluating is an important and interesting management which problem in financial analysis. *Francesca et al* proposed a time hazard model for a population of loans involves different probability of default considering conjointly the explanatory variables and the time when the default occurs. Good borrowers for which the risk of default is the lowest and bad borrowers for which this risk is the highest.

Purohit et al proposed that checks the applicability of the new integrated model on a sample data taken from Indian bank. This is an integrated combination model based on decision tree,

Support vector machine; logistic regression and Radial basis neural network and compares the effectiveness of these techniques for approval of credit. The possibility of connecting unsupervised and supervised techniques for credit risk evaluation was proposed by *Zakrzewska et al.* These technique presented building of different rules for different group of customers and in this approach, each credit applicant is assigned to the most similar group of clients from the training data set and credit risk is evaluated by applying the appropriate rules

for the group. *Bhasin et al* proposed to extract important information from existing data and enables better decision making in banks. Data warehousing is used to combine various data from databases into an acceptable format so that the data can be mined. The tools of data mining are analyzed in data warehousing rule selection mechanism is introduced by *İkizler et al.* This new method has been applied for learning interesting rules for the evaluation of bank loan application. A decision tree classifier is used in generating the rules of the domain. *Nassali et al* proposed a new loan assessment system and developed prototype software for this system. According to this, the effective use of this system will make a positive impact on the quality of the decisions made. This will save the time from the application of loan. So assist in reducing the size of labor and the number of bad debts. *Jacobson et al* proposed a bivariate probit model to investigate the implications of bank lending policy is applied. A value at risk measure is derived for the sample portfolio of loans and show how this can enable financial institutions to evaluate alternative lending policies on the basis of their implied credit risk and loss rate.

Karaolis et al proposed a method to develop a data mining system for the assessment of heart related risk. Data mining analysis is carried out by using decision tree. *Anbarasi et al* proposed an acurate prediction is done by feature subset selection of attributes. The attributes are reduced using genetic algorithm. Classification is done based on three classifiers like Naïve Bayes, Decision tree and classification via clustering to predict the diagnosis of patients with the same accuracy as obtained before the reduction of attributes. The method of selecting or choosing the best attribute based on information entropy was proposed by *Du et al*. This paper shows the procedure for selecting the decision attribute in detail and finally it points out the developing tends of decision tree.

Karaolis et al proposed the Assessment of the Risk Factors of Coronary Heart Disease (CHD) is done based on data mining. In this method the attributes are selected based on two bases: non-modifiable and modifiable. The attributes that occurred after the event of CHD are also considered like: smoking after the event, systolic blood pressure, diastolic blood pressure, total cholesterol, high density lipoprotein, low-density lipoprotein, triglycerides, and glucose. Since this existing method can be utilized only in medical applications, a new method (ERPCA) is used in the proposed method which can be used in bank applications method aids the bank by making efficient risk assessment of whether a loan can be sanctioned to a particular customer or not, than the existing methods. The experimental results shows that the proposed method has greater accuracy in classification of customers as good and bad based on the risk factors. In this method bank database (customer details) are used as inputs in which different attributes like age, sex, marital status, occupation, minimum age, maximum age, maximum experience, annual income, net profit, other loan s(if any loans the customer received from other banks) etc. of a customer are considered for further processing.

ERPCA Method

This algorithm evaluates the risk of multidimensional data based on risk prediction clustering algorithm.Credit scoring is defined as a statistical method that is used to predict the probability that a loan applicant will default or become delinquent.

Credit scoring helps to increase the speed and consistency of the loan application process and allows the automation of the lending process.Risk assessment is one of the existing problems in the bank sector. The decision for the credit sanction to a customer should be evaluated properly so that, it may not lead to loss for the Bank. The existing method (ERPCA) aids the banking sector to make the evaluation for loan sanction in an enhanced manner. Rules are formed for each loan type like (personal loan, bike loan, car loan, house loan, business loan). Associative clustering algorithm (ERPCA) is used to mine the clusters from massive and high dimensional numerical databases[17].

A group of data elements can belong to more than one cluster, which is associated with each element is a set of membership levels.Using ERPCA algorithm, three vectors can be taken into consideration.The centroid and coefficient of classified data is computed and the obtained result is compared with three initialized vectors. The variables L1, L2, M1, M2, H quoted in this algorithm takes the value of 0 and 25 for low, 26 and 50 for medium and greater than 50 for high. Based on these three vectors, the data are clustered.

III. Methodology for Multi-Dimensional Rule Prediction using K-means Clustering

Risk prediction is an important issue in banking sector. In order to avoid credit loss in bank, credit sanction to a customer has to be decided effectively. The proposed method aids the banking sector to evaluate the loan particulars in an effective manner.

In this method, customer details those who applied for loan are collected and remove the unnecessary information by feature extraction process. Association rules are generated for each loan type like personal loan, home loan, car loan etc., Based on the rules, risk assessment is performed by two levels such as primary and secondary. Finally, loan applicants are grouped based on the prediction as accepted or rejected loan applicants by k-means clustering algorithm. The overall flow of the proposed system is as below.

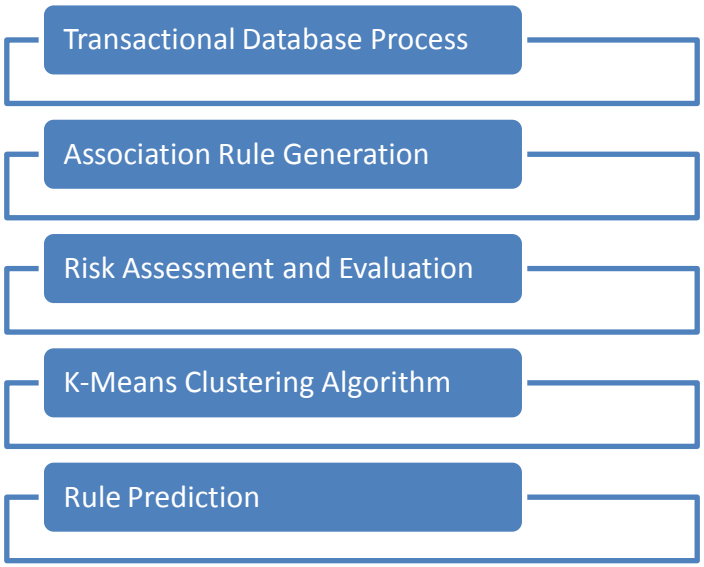

Figure 1 Overall Flow of Proposed system

K-Means Clustering

Clustering groups the similar set of objects. K-means clustering is applied for mining clusters efficiently from high voluminous datasets. ***k*-means clustering** is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. *k*-means clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. *k*-means clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes.

The algorithm has a loose relationship to the *k*-nearest neighbor classifier, a popular machine learning technique for classification that is often confused with *k*-means because of the *k* in the name. One can apply the 1-nearest neighbor classifier on the cluster centers obtained by *k*-means to classify new data into the existing clusters. This is known as nearest centroid classifier.

Clustering indicates the strength of association between data element and particular customer.

Using proposed algorithm, three rules can be taken into consideration such as Low, Medium, and high. Based on these criteria's, risk assessed data's are clustered. Mean value is calculated and obtained result is compared with three criteria's. The variables L_1, L_2, M_1, M_2, H denoted in the algorithm takes the value of 0 to 25, 26 -50 and greater than 50. Based on these criteria's, similar data's are clustered and stored in the dataset.

Proposed Algorithm

Input: Cluster(t), L_1,L_2,M_1,M_2, H,E

Begin

Clusters $t_k(x)$ = coefficients

Repeat until when $t_k(x) < \in$

Centre for each X

for each x $C_k=1/m \Sigma^{m}_{j=1} t_{ij}$

if $C_k \geq L_1$, && $C_k \geq L_2$ then

$CL+= C_k$

else if $C_k \geq M_1$ && $C_k \leq M_2$ then

$C_M+= C_k$

else if(C_k >H) then

$C_H+= C_k$

end if

end for

Collect all clusters C_L,C_M,C_H

Rule prediction

To sanction loan, threshold value is initial and predicted the risk value is based on threshold limit. Loan approval and loan rejection list are classified using this threshold limit and then the customers are clustered separately for efficient processing.

IV. Comparative Analysis

The proposed method is compared with existing ERPCA technique. The existing method makes risk assessment by centroid using association clustering algorithm. But still the risk percentage accuracy is not good and efficient. So method overcomes these drawbacks by proposed model using k-means clustering method. Experimental result shows that the proposed framework evaluates the risk in the given set with better accuracy and consumes less time than existing techniques.

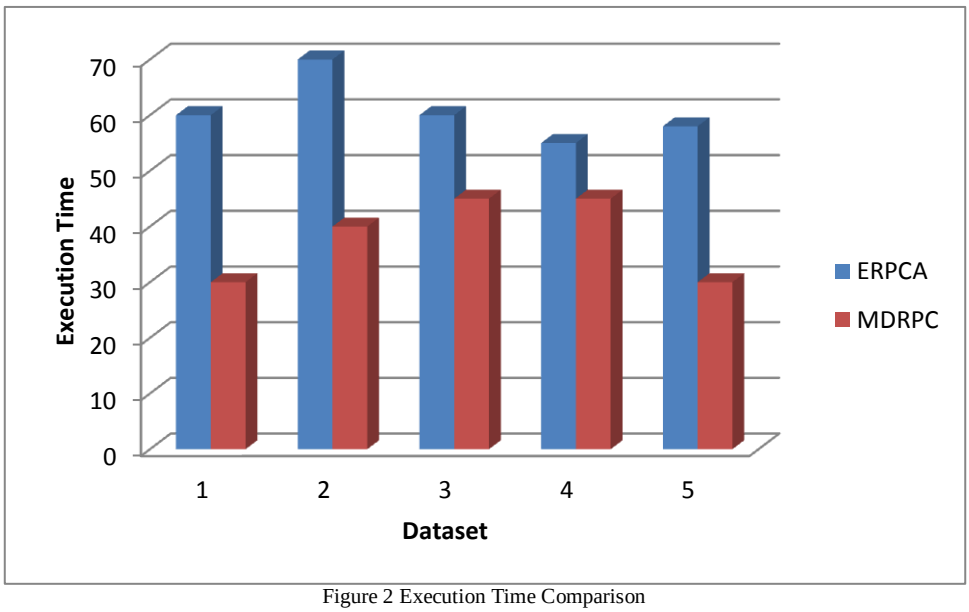

Figure 2 Execution Time Comparison

V. Conclusion

Risk Assessment and Evaluation are the difficult tasks in finance sectors. This paper proposed a new framework which is integrated by k-means clustering algorithm, association rule mining and rule prediction method. Clustering techniques separate the customer status such as good and bad based on predefined criteria's which is fixed by bank. Here duplications are avoided by using Association Rule. It is clearly projected that the proposed work provides better accuracy than existing method.

References

1. G. Francesca, "A Discrete-Time Hazard Model for Loans: Some Evidence from Italian Banking System," American Journal of Applied Sciences, vol. 9, p. 1337, 2012.
2. S. Purohit and A. Kulkarni, "Credit evaluation model of loan proposals for Indian Banks," in Information and Communication Technologies (WICT), 2011 World Congress on, 2011, pp. 868-873.
3. D. Zakrzewska, "On integrating unsupervised and supervised classification for credit risk evaluation," Information Technology and Control, vol. 36, pp. 98-102, 2007.
4. M. L. Bhasin, "Data Mining: A Competitive Tool in the Banking and Retail Industries," Banking and finance, vol. 588, 2006.
5. N. İkizler and H. A. Guvenir, "Mining interesting rules in bank loans data," in Proceedings of the Tenth Turkish Symposium on Artificial Intelligence and Neural Networks, 2001.
6. J. Nassali, "A Loan Assessment System for Centenary Rural Development Bank," 2005.
7. T. Jacobson and K. Roszbach, "Bank lending policy, credit scoring and value-at-risk," Journal of banking & finance, vol. 27, pp. 615-633, 2003.
8. G. Kabir, I. Jahan, M. H. Chisty, and M. A. A. Hasin, "Credit Risk Assessment and Evaluation System for Industrial Project."
9. B. Bodla and R. Verma, "Credit Risk Management Framework at Banks in India," ICFAI Journal of Bank Management, Feb2009, vol.8, pp. 47-72, 2009.
10. R. Raghavan, "Risk Management in Banks," CHARTERED ACCOUNTANT-NEW DELHI-, vol. 51, pp. 841-851, 2003.
11. M. A. Karaolis, J. A. Moutiris, D. Hadjipanayi, and C. S. Pattichis, "Assessment of the risk factors of coronary heart events based on data mining with decision trees," Information Technology in Biomedicine, IEEE Transactions on, vol. 14, pp. 559-566, 2010.
12. M. Anbarasi, E. Anupriya, and N. Iyengar, "Enhanced prediction of heart disease with feature subset selection using genetic algorithm," International Journal of Engineering Science and Technology, vol. 2, pp. 5370-5376, 2010.
13. M. Du, S. M. Wang, and G. Gong, "Research on decision tree algorithm based on information entropy," Advanced Materials Research,vol. 267, pp. 732-737, 2011.
14. X. Liu and X. Zhu, "Study on the Evaluation System of Individual Credit Risk in commercial banks based on data mining," in Communication Systems, Networks and Applications (ICCSNA), 2010 Second International Conference on, 2010, pp. 308-311.

15. B. Azhagusundari and A. S. Thanamani, "Feature selection based on information gain," International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN, pp. 2278-3075.

16. M. Lopez, J. Luna, C. Romero, and S. Ventura, "Classification via clustering for predicting final marks based on student participation in forums," Educational Data Mining Proceedings, 2012.

17. K.Kala, Dr. E.Ramaraj "ERPCA: A Novel Approach for Risk Evaluation of Multidimensional Risk Prediction Clustering Algorithm" ,International Journal of computer science and Engineering, ISSN : 0975-3397 Vol. 5 No. 10 Oct 2013.

