Volume 6, Issue 2, February 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Clustering Loan Applicants based on Risk Percentage using K-Means Clustering Techniques

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

Abstract: ***Data mining techniques uses some key ideas for data classification and prediction. Clustering techniques is used to place data items in to similar groups without prior knowledge of group definitions. Clustering provides efficient decision making by grouping large voluminous datasets in bank. Risk assessment is an important task of bank, as the increase and decrease of credit limits in bank depends largely to evaluate the risk properly. The key problem consists of identifying good and bad customer's status those who applied for loan. An improvised risk evaluation of Multi-dimensional Risk prediction clustering Algorithm is implemented to determine the good and bad loan applicants whether they are applicable or not. In order to increase the accuracy of risk, risk assessment is performed in primary and secondary levels. Hence for avoiding Redundancy, Association Rule is integrated. This method allows for finding the risk percentage to determine whether loan can be sanctioned to a customer or not. Finally it is proven that proposed method predicts the better accuracy and consumes less time than existing method.***

Keyword: Risk Prediction, Clustering, Redundancy, Data Mining, Feature Extraction

I. INTRODUCTION

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.

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 has 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.

Credit Scoring

Credit scoring is defined as a statistical method which is used to predict the probability. This helps to determine whether credit should be granted or not to a borrower. Credit Scoring can also be defined as a systematic method for evaluating credit risk that provides consistent analysis of the factors that have been determined to affect the level of risk. Credit scoring helps to increase the speed and consistency of the loan application process and allows the automation of the lending process. Also, it greatly reduces the need for human intervention on credit evaluation and the cost of delivering credit.

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

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. 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 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. Figure 1 depicts the attributes used in the existing method.

III. IMPLEMENTATION OF PROPOSED METHODOLOGY

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.

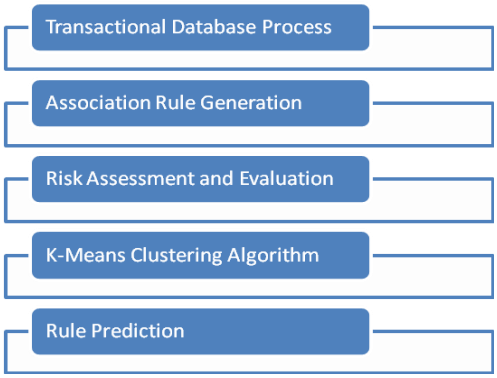


Figure 1 Overall Flow of Proposed system

Here, each customer who needs loan has to provide their personal details, Income details, loan details, occupation details etc., These details are stored in the bank database for further processing. Customer details are prepared in suitable manner for data mining. The bank database contains the following attributes.

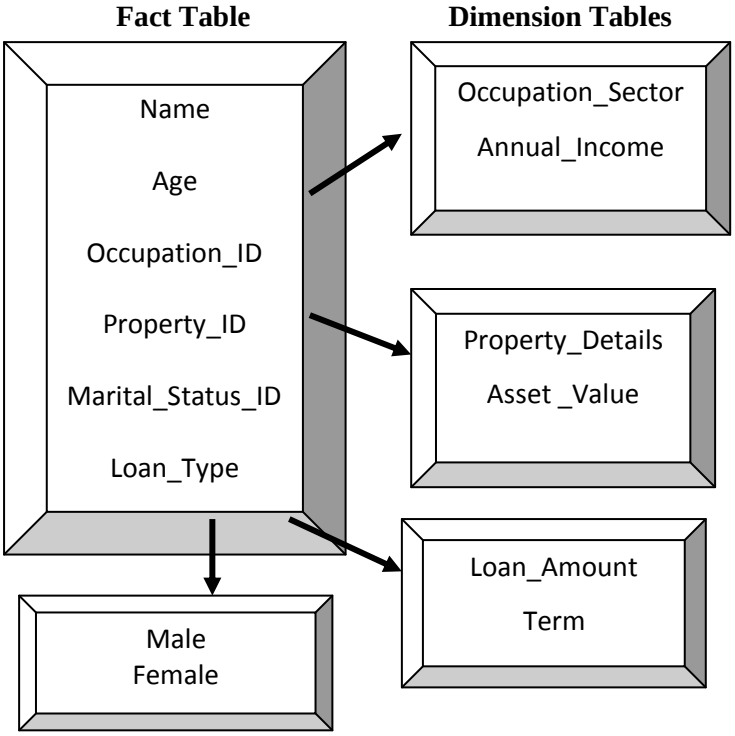


Figure 2 Fact and Dimension Table for Loan Applicants

Customer details are collected in the database and segmented based on loan type. Then the valid attributes are selected using feature extraction process.

Transactional Dataset Processing

Customer details are extracted and stored in the bank database. In order to avoid time consuming & memory complexity, unnecessary attributes are filtered from transactional database. In the bank dataset, loan type, loan_amt, occupation, sector, age, experience, annual_income, term, Net profits are considered as valid attributes form this process. This reduces the degree of attribute size. The main goal of the process is to identify the minimum set of attributes and reduces the complexity. Feature extraction is performed for filtering irrelevant attributes.

Association Rule generation and Risk Assessment

Each bank has different criterias which has to be satisfied by the customer to avail loan. To receive a specified loan, customer has to satisfy particular touch stones like age, Annual_Income, Loan Amt, Term. Based on the loan type, applicants of the attributes are altered. Rule list are generated by the bank as shown in below table.

Loan_Type	Sector	Age Limit	Annual Income	Amount	Term
Personal	Govt	21-65	>5L	6L	5
Personal	Self_Employed	30-60	>10L	3L	3
Housing	Govt	21-55	>10L	75% of Value	10
Housing	Self_Employed	25-65	>15L	30% of Value	7
Car Loan	Govt	21-65	>3L	80% of Value	5
Car Loan	Private	30-60	>3L	75% of Value	5
Car Loan	Self_Employed	21-60	>10L	70% of Value	3

Bank officers should be capable of measuring the risk for loan. Risk assessment is performed by measuring some attributes. Here the risks are classified into two categories such a primary and secondary. Major criteria portions are considered as primary such as loan_amt, net_profit or Annual_income. Secondary risk is calculated by using age_limit, experience & occupation. Based on the predicted rules, values are assigned for the corresponding attributes. For example if the customer satisfies the criteria such as age_limit is in between the criteria then the value is assigned 1 else 0. Primary and secondary levels are identified by calculating the average of corresponding attributes. Risk percentage is calculated using the following equation

$$Risk\_Percentage = (1 - risk) * 100$$

where

$$Risk = (0.6 * Primary\_Risk) + (0.4 * Secondary\_Risk)$$

Risk percentage is calculated to evaluate whether loan can be sanctioned to an applicant or not. Then the customers are classified based on the risk percentage obtained.

In the proposed algorithm, user has to specify their loan particular and personal details. User Information and rule list generated by the bank are given as input. Threshold value should be initialized. The loan type is compared between

customer and rule list dataset. If it agrees and then proceed to check all the criteria's assigned by rule list than finally risk value is calculated which is compared with threshold limit. If the risk percentage is less then threshold_ limit then the loan is sanctioned otherwise not sanctioned.

K-Means Clustering

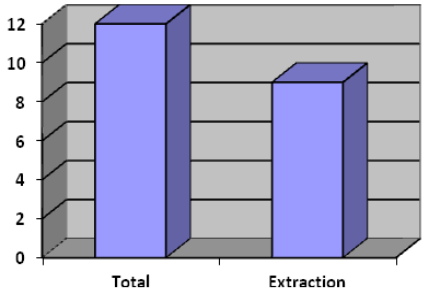
Clustering groups the similar set of objects. K-means clustering is applied for mining clusters efficiently from high voluminous datasets. Clustering indicates the strength of association between data element and particular customer. Using proposed algorithm, three rules can be taken into consideration 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.

Rule prediction:

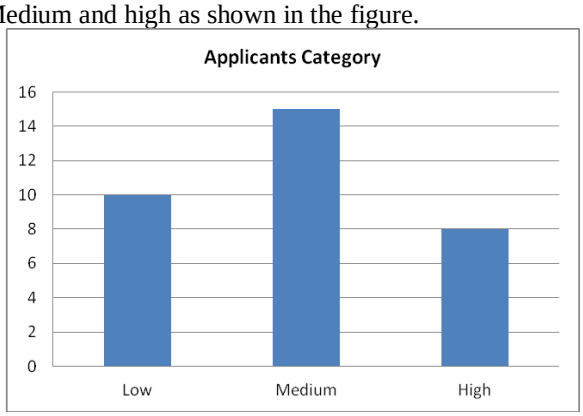
To sanction loan, threshold value is initial and predicted the risk value is based on threshold limit. Loan approval and loan rejection list are classified using this threshold limit and then the customers are clustered separately for efficient processing.

IV. EXPERIMENTAL RESULTS

To evaluate the effectiveness of the proposed work, performance evaluation is carried out. 1000 loan applicants are extracted with their personal details. Initially it consists of 12 attributes. Loan applicants are segmented based on loan type. Among 12 attributes, during feature extraction 3 attributes are selected for primary risk and 6 attributes are selected for Secondary risk level Consideration. The following figure depicts the feature extraction process.



After extraction process, rules are predicted bank proposed different rules based on loan type and customer occupation.Primary risk considers. Annual _Income, Loan_amount, Occupation and term than secondary risk considers age, experience etc., Using these attributes, risk percentage is calculated and threshold value is fixed by the bank. Customers risk percentage is more than threshold values are considered as risking customer. Based on risk percentage, customers are classified as Low, Medium and high as shown in the figure.



V. CONCLUSION

Risk Assessment & Prediction is a critical task in banking industry. This paper proposes a framework for risk evaluation based on k-means clustering techniques. Customer data are extracted and the relevant attributes are selected using Information gain theory. Rule prediction is performed for each loan type based on the predefined criteria's. Accepted and rejected applicants are considered as "applicable" and "Non-applicable" credits accordingly experimental results shown that the proposed method predicts better accuracy and consumes less time than the 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

