A Customized Approach for Risk Evaluation and Prediction based on Data Mining Technique

K. Kala
Research Scholar,
Manonmaniam Sundaranar University,
Tirunelveli

***Abstract*—Risk assessment is important task of Banks, as the failure and success of the Bank depends largely on banks' ability to evaluate the credit risk properly. Decisions regarding credits granting are one of the most crucial issues in an everyday banks policy. The credits that are well allocated may become one of the biggest sources of profits and any mistakes in decision may lead to a loss. The key problem consists of distinguishing, salubrious (good) and delinquent (bad) credit applicants. The main objective in credit risk evaluation consists of building classification rules that assign bank customers as good or bad payers. The proposed method, customized approach for risk evaluation (CARE) lays down a risk evaluation process to determine the good and bad loan applicants using data mining technique. The attributes of the customers are selected and features are extracted for efficacy. Rules prediction is done for each type of loans to avoid redundancy. The risk assessment consists of two levels, primary and secondary levels. This method allows for finding percentage of risk to determine whether loan can be sanctioned to a customer or not. C4.5 algorithm is used to classify the risk levels as low, medium and high. The system is tested by generating loan applicants on own and the results show that our proposed method is efficient over the existing methods.**

***Index Terms*— Attributes, credit granting, credit risk, percentage of risk, and risk assessment.**

I. INTRODUCTION

Due to high competition in the business field, it is essential to consider the customer relationship management of the enterprise. Here analyze the massive volume of customer data and classify them based on the customer behaviors and prediction. Customer relationship management is mainly used in sales forecasting and banking areas. Data mining provides the technology to analyze mass volume of data and detect hidden patterns in data 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.

Data mining is the extraction of required data or information from large databases. The key ideas are to use data mining techniques to classify the customer data according o the posterior probability. Here the Data mining concept is used to perform the classification and prediction of loan.

With the continuous development and changing in the credit industry, credit products play a more and more 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 huge amount of losses. It is an even more important task today as financial institutions have been experiencing serious challenges and competition during the past decade. When considering the case regarding the application for a large loan, such as a construction loan, the lender tends to use the direct and individual scrutiny by a loan officer or even by a committee. The extent to which a borrower uses the credit facility, greatly impacts the repayment ability and performance of the firm, which then affects the lending institutions. It is therefore, of paramount concern to lenders to limit potential default risks, screening the customer's financial history and financial background. Banks should control credit management thoroughly. Sanctioning of loan requires the use of huge data and substantial processing time. Before sanctioning/ 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. Sometimes with flooded work load, and lack of new technologies, the decisions of sanctioning loans may become wrong and resulted in credit defaults. An intelligent information system that is based on C4.5 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 that is used to predict the probability that a loan applicant will default or become delinquent. 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 cause or affect the level of risk. The objective of credit scoring is to help credit providers quantify and manage the financial risk involved in providing credit, so that they can make better leading decisions quickly and more objectively. Credit Scoring has many benefits that accrue not only to the lenders but also to borrowers. 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. Credit scores can help financial institutions determine the interest rate that they should charge their customers and to price portfolios. The methodology for constructing credit scoring models generally involve following process. Credit scoring task can be divided into two different types: The first type is application scoring, where the task is to classify the credit applicants into accepted or rejected groups. The data used for modeling generally consists of financial information and demographic information about the loan applicant. The second type of tasks deals with the existing customers and along with other payment history information is used here. This is distinguished from the first type because this takes into account the customer's payment pattern on the loan and the task is called behavioral scoring. Then a deductive credit scoring system awards points (weights) to particular relevant attributes of the credit customers. The weighed value of attributes is aggregated to total score. The relevant attributes and their weights are determined by the credit decision makers based on their experiences.

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 (CARE) are given in section 3. Empirical study is 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 evaluation is an important and interesting management problem in financial analysis. The authors of [1] Uses the theory of artificial neural networks and business rules to correctly determine whether a customer is default or not. The Feed-forward back propagation neural network is used to predict the credit default. Seema et al. [2] proposed a paper that checks the applicability of one of the new integrated model on a sample data taken from Indian bank. This is an integrated combination model based on the techniques of 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 is investigated in [3]. The technique presented allows 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. According to [4] Data mining is a tool used to extract important information from existing data and enables better decision making in banks. They use data warehousing to combine various data from databases into an acceptable format so that the data can be mined. The concepts and tools of data mining are analyzed in this. Based on [17] the data preprocessing techniques like data reduction and data cleaning can be applied for data preparation and dates were converted into numerical form. A data model is generated and classified using Naïve Bayesian algorithm and placed appropriately based on posterior probability and based upon this the percentage of loan sanction can be predicted. Rule interestingness measures are discussed ad a new rule selection mechanism is introduced in [5]. This new method has been applied for learning interesting rules for the evaluation of bank loan application. C4.5, a decision tree classifier, is used in generating the rules of the domain. Nassali [6] 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 right from the application of loan to the sanction of loan. This will also assist in reducing the size of labor and the number of bad debts. As per [7] a bivariate probit model to investigate the implications of bank lending policy is applied. They derive a value at risk measure 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. Kabir et al. [8] has adopted a standardized approach in the form of credit risk grading (CRG) system to assist the improvement in the banking sector. This whole model is divided into six risk components and each risk is again divided based on some criteria which are considered as crucial risk determinants and further criteria are scored against specific parameters in order to assess the final grading score. In [9] an attempt has been made to study the Credit Risk Management Framework of scheduled commercial banks operating in India. The effectiveness of Risk management in banks depends on efficient Management Information system, Computerization and net working of the branch activities [10].

The data warehousing solution should effectively interface with the transaction system like risks systems and core banking lists to collate data. Karaolis et.al [11] proposed a method to develop a data mining system for the assessment of heart related risk. Data mining analysis is carried out using decision tree C4.5 algorithm. According to [12]C4.5 is one of the most popular algorithms for rule based classification. This algorithm has many empirical features such as missing value handling, continuous number categorization, etc. Most related attributes should be selected from a dataset to perform higher accuracy using C4.5. Entropy of Information theory is measured to identify the central attribute of the dataset. Decision tree is an important method in Data mining. The paper [13] discusses the method of selecting or choosing the best attribute based on information entropy. This paper shows the procedure for selecting the decision attribute in detail and finally it points out the developing tends of decision tree. An Individual Credit Risk Evaluation System (ICRES) using data mining technology, with the aid of 'feed forward control' in management theory is proposed in the paper [14]The information gain method is used to screen the alternative indicators that have greater impact on the classification prediction. In paper [15] the authors proposed an algorithm based on information gain and discernibility matrix to reduce the attributes. The reduction of attributes is one of the important processes for knowledge gaining. The classification of multidimensional and larger datasets, leads to wrong results. The features are mostly inconsistent and redundant which affects the classification. The method proposed in this paper overcomes all these disadvantages. Ahmad Nadali et al. [16] proposed a hybrid method for evaluating credit risk of bank customers. Significant financial ratios are extracted from the balance sheet and kolmogorove-Smirnov test is used to specify kind of financial ratios distribution. Meaningful variables are obtained by using T test and effective ones are determined using DEMATEL method. Credit risk is finally assessed using Fuzzy expert system.

III. PROPOSED METHOD

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 proposed method (CARE) aids the banking sector to make the evaluation for loan sanction in an enhanced manner. The overall flow of the proposed work is shown in figure (1).

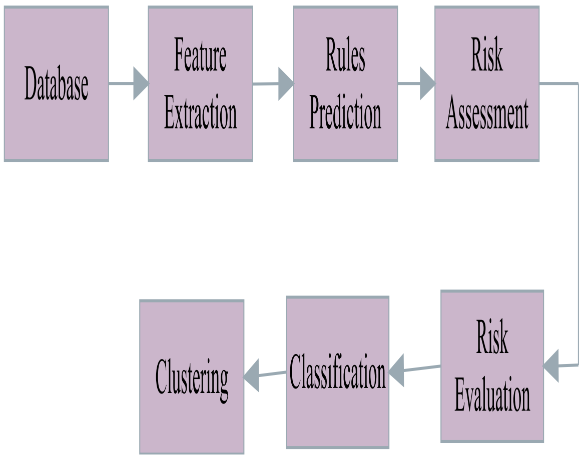

Figure 1: Overall Flow of the proposed work

Here, each bank customer who needs loan has to provide their personal details, income details and loan details etc to the bank. These details are stored in the database for further access. The details of the customers are to be prepared and cleaned in such a manner suitable for data mining. The data preparation process includes collection of details from the customers and converting into a format suitable for the data mining. The data cleaning process includes cleaning of data by filling the missing values, extraction of duplications, removing outliers and resolving inconsistencies.

The dataset contains attributes like Age, Sector, Years of Experience, Property, marital status, Nominee, Other loans, Amount, Term, Net profit, Asset value. The details of the applicants are collected in the database and then segmented based on loan type. Then the valuable attributes are selected using Feature extraction.

3.1 Segmenting Customers

In Banking, customer segmentation allows reduced exposure to risk assessment. It also allows personalized services according to client interests and matching campaigns to customers. A classifier model is essential in customer segmentation area. Using this technique customers are classified based on loan type so that all the customers seeking for a particular loan are combined together.

3.2 Feature Extraction

Data sets for analysis may contain many attributes, which may also contain irrelevant data to the mining task. Though it is possible for a domain expert to pick out the essential attribute, it may consume time and make the task difficult. Keeping of irrelevant attributes leads to confusion in mining process and also increases the data size. Thus the attributes has to be reduced to decrease the size. The goal of this process is to find a minimum set of attributes to help in easier understanding of data and to reduce the computational complexity. Here, Feature extraction is done by calculating Information gain.

3.2.1 Information Gain

Information theory is widely used in data mining. In this, entropy measures the uncertainty among random variables in the database. Claude E. Shannon has introduced the idea of entropy of random variables. Entropy provides the long term behavior of random process that is very useful to analyze data. The behavior of the random process is a key factor for developing the coding for information theory. Entropy is a measurement of average uncertainty of collection of data when we do not know the outcome of an information source. Entropy is the measurement of how much information we do not have. Info gain of an attribute is used to select the best splitting criterion attribute. The highest value Info Gain is used to build the decision tree.

$$Infogain\ (A) = Info(D) - Info_A(D) \qquad (1)$$

Where A is the attribute investigated.

$$Info(D) = -\sum_{i=1}^{m} p_i \ log_2(p_i) \qquad (2)$$

Where
$p_i = probability\ (class\ t\ in\ dataset\ D)$

$m = number\ of\ class\ values$

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} info(I \qquad (3)$$

```
Algorithm: Risk calculation
Input: user info (u_list) , rules list(r_list)
Output: Risk percentage
Begin
    Initialize threshold value
    For Each loan type in r_list
    If (u_list.loan type = r_list.loan type)
    Goto step 1
     else end
 Step 1: initialize s =0
        if (u_list.occupation = r_list.occupation && u_list.sector=r_list.sector)
          if (u_list age>= r_list.min age && u_list age<= max age )
              s+2
             if (u_list.min service>= r_list.min service)
                 s +1
                  if (u_list.annual income>= r_list.annual income)
                      s+1
                       if (u_list.required amount<= r_list.max amount)
                           s+1
                           if (u_list.term<= r_list.term)
                              s+1
             Goto step 2
        else end
Step 2:          x =s/n *100                   /* calculate risk */
                  Risk = (1-x)
                  if(Risk <= threshold)
                  return true
                  else
                  return false
```

|D| = total number of observations in dataset D
= all attribute values.

Info Gain is calculated using the equation (1). Here, first the entropy for all attributes in the dataset are calculated using eqn. (2) and then the entropy for individual attributes are calculated using the eqn. (3). Entropy for individual attributes is subtracted with total entropy to get the info gain. The attribute with highest info gain is used further for the data mining process. Thus the feature extraction process is done to select the valuable attributes.

3.3 *Algorithm for Rules prediction and Risk evaluation*

Loan type	Occupation	sector	Min age	Max age	Min experience	Annual income	Amount	Term
personal	Selfemployed	Business	25	65	3	EMI*24	300000	3
Housing	Salary based	Govt.	21	45	1	EMI*24	80%ofvalue	10
Business	salary	private	21	48	1	EMI*24	25%ofnet profit	7
Car loan	Selfemployed	Business	25	55	3	EMI*24	80%ofvalue	5
Bikeloan	Salary	Govt.	21	45	2	EMI*24	70%ofvalue	3

Table 1: Rules Predicted

3.3 *Rules Prediction*

Each bank has different rules criteria that have to be satisfied by the customer to get the loan. Apart from the rules created earlier, new rules can also be introduced in this method. To receive a particular loan, customer has to satisfy particular touchstones like Minimum age, Maximum age, Minimum service, ROI, Annual income, Maximum amount and Maximum years as shown in the table 1. The appreciates of these attributes is altered based on the loan type and user information enforced by the applicant.

3.3.1 *Risk Assessment and Evaluation*

In order to price a loan a lending officer should be capable of measuring the risk attached to the loan. Risk assessment is done by measuring certain attributes. Here, the risks are separated into two categories: Level I and Level II type risk. They can be considered as primary and secondary risks. Primary risk is calculated by considering the three attributes Amount, Term and Netprofit. Secondary risk is calculated by using the attribute Minimum Age, Maximum Age, Minimum Exp. These attributes are selected based on the values obtained from Info Gain.

Based on the predicted rules, values are assigned for the attributes. For example: if the minimum age of the customer satisfies the rules predicted for the loan he applies, then the value is assigned 1 else 0. Level I and II risks are found by calculating the average of corresponding attributes.

Using Level I and Level II risk, % of risk is calculated. % of risk is calculated using the equation (4). In this, greater weight age is given to Level I risk.

$$\%\ of\ risk = (1 - risk)*100 \qquad (4)$$

Where

$$Risk = (0.8*\ Level\ I\ risk) + (0.2*\ Level\ II\ risk) \qquad (5)$$

% of risk is calculated for all the customers applied for loan, to evaluate whether loan can be sanctioned to a particular customer or not. Then the customers are classified, based on the % of risk obtained.

According to the algorithm 3.3, the user has to specify his loan type and all his personal details as inquired by the bank. Peculiar rules are framed by the bank and preserved as list. In the algorithm, user information and rules list are given as input. User info is considered as u_list and rules list are considered as r_list. Threshold value should be initialized. The loan type specified by the user is compared with the loan type in the rules list. If it agrees, then it goes for step 1 where comparison is made for sector and occupation. If these both criteria also agree, then all other six attributes of level I and level II are compared and value of s is incremented for all true comparisons and percentage of risk is calculated. When the condition does not satisfy, the procedure ends. The risk calculated is compared with the threshold value. When the percentage of risk is less than the threshold value then the loan is sanctioned, else rejected.

3.5 Classification

In this process, well known, C4.5 algorithm is used. It is based on ID3 decision tree induction algorithm enhanced with improvements concerning with missing values, numeric attributes, and noisy data and generating rules from trees. Rules prediction are done for every segmented data formed before based on loan type. C4.5 constructs a very big tree by considering all attribute values and finalizes the decision rule by pruning. Decision tree is used to classify a case, i.e. to assign a class value to a case based on the values of the case. A path from the root to a leaf of the decision tree can be followed based on the attribute values of the case. The C4.5 algorithm constructs the decision tree with a divide and conquers methodology. Each node in a tree is related with set of cases. Cases are assigned weights to take into account of unknown attribute values. C4.5 can produce both decision tree and rule sets and can construct a tree for the purpose of improving the prediction accuracy.

Here, C4.5 is used to frame rule sets to classify the risk levels. The risk levels are classified based on the % of risk values. They are classified based on the three vectors low, medium and high. The values below 25 is assigned as low, 25- 45 is medium and above 45 is high. Now, the classified data are clustered based on these three vectors.

For the loan sanction, a threshold value of 35% of risk is set. Based on this value, lender can decide whether to sanction loan or not i.e. if the % of risk for a customer is greater than 35%, the application is rejected, else loan is sanctioned. Loan approval list and Loan rejection list are classified using this threshold value. Then the loan approved customers and loan rejected customers are clustered separately for efficient retrieval.

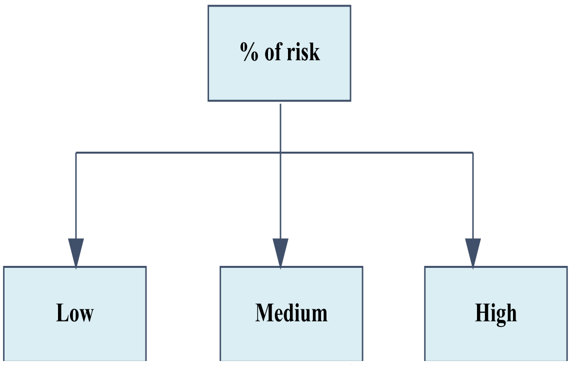

Figure 3: Three levels of Risk

IV. EXPERIMENTAL RESULT

To evaluate the effectiveness of the proposed Risk evaluation technique, a series of experiment is percolated thereby performance validation is carried out. To start-off with this method the experimental dataset are generated on own. 2000 loan applicants are generated with their personal details. Initially these details consist of 15 attributes. The loan applicants are segmented based on the loan type they seek for (example: bike loan or house loan etc.). The next process is feature selection. Feature selection is done to extract the attributes. This is normally carried out to eliminate redundant and irrelevant features that are extracted. Here, author estimates the entropy for the feature selection. Having entropy values determined the mutual information among the features and the targets are determined. This information is used to estimate and measure how a random variable is able to describe and impact on other variable. Among the 15 attributes, 13 attributes are selected by feature extraction process. Figure 4.depicts the number of features present before and after feature extraction process.

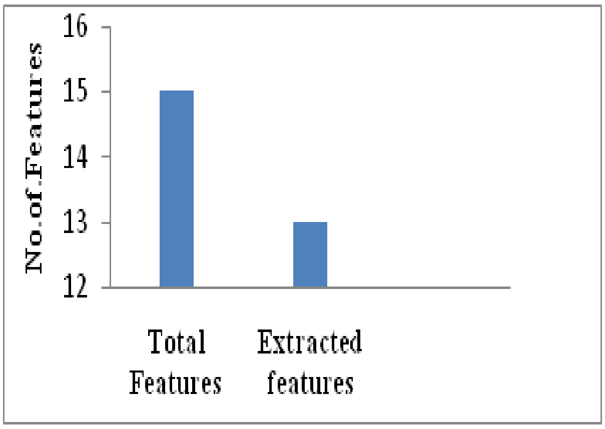

Figure 4: Features present before and after Extraction process.

After the feature extraction process, rules are predicted. Bank proposes different rules for the loan applicants, based on the loan type. Then, Risk Assessment is done in two levels, primary and secondary. The primary risk analysis considers certain attributes such as Amount, Term and Net profit. Secondary level of risk considers attributes such as Minimum Age, Maximum Age and Minimum experience. Using these attributes % of risk is calculated. Threshold value is fixed, so that the customers with % of risk more than this threshold value are considered as risky customer and rejected. If else, loan is sanctioned for the customer. Based on the % of risk value calculated, customers are classified as Low, medium and high as shown in figure 5.

For the classification of dataset, C4.5 algorithm is used. The performance analysis of C4.5 algorithm is compared with ID3 algorithm. The limitation of ID3 is that it is very sensitive to features with large numbers of values. To overcome this, C4.5 uses a metric called “information gain,” which is defined by subtracting conditional entropy from the base entropy that is Gain.

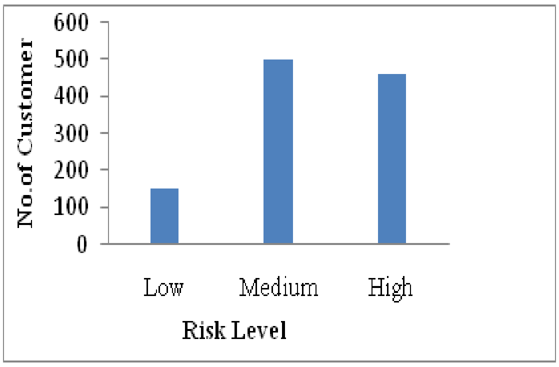

Figure 5: Risk levels obtained after classification

C4.5 algorithm is chosen in this method as it has, Greater accuracy, less memory usage, less search time .and less time taken to build model. From the graph, we come to know that C4.5 has a greater accuracy and consumes less time.

The comparison of accuracy and time taken by C4.5 and ID3 are depicted in the figure 6 and 7

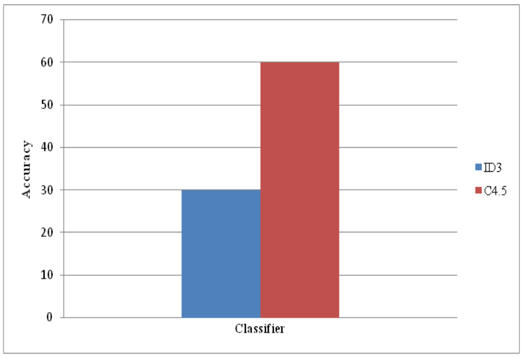

Figure 6: Accuracy comparison for ID3 and C4.5

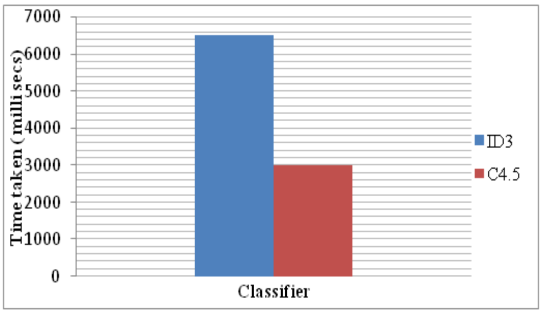

Figure 7: Time utilization comparison for ID3 and C4.5

Finally, based on the percentage of risk value, loan applicants are separated as good credits and bad credits as shown in the figure 8.

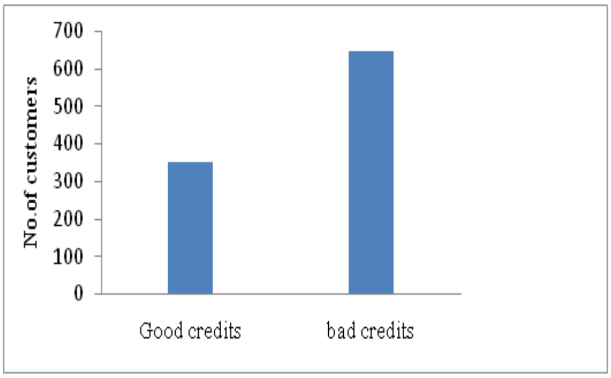

Figure 8: Illustration of good and bad credits

Comparison with the existing system

The proposed framework (CARE) is compared with an existing technique proposed in [12]. The existing method makes risk assessment only for cancer prediction. There are no methods proposed until, for the evaluation of risk in bank credit sanction. But the proposed work overcomes this disfavor and establishes the risk assessment in banks. Experimental results show that the proposed CARE framework evaluates the risk in the given set of documents effectively than the existing techniques. The authors of this paper have planned to apply the framework to support versatile applications.

V. CONCLUSION

Risk Assessment is the crucial task in the Banking industry**.** This paper proposes a framework (CARE) for risk evaluation, where mass volume of customer data are engendered and risk assessment plus evaluation is done based on the Data mining technique. The customer data are extracted for feature selection of the valuable attributes. The attributes are selected using Information gain theory. Rules prediction is done for each loan type. Risk assessment is performed in two levels, primary and secondary namely. Each risk levels consist of three attributes to be evaluated.C4.5 algorithm is used to classify the risk levels as low, medium and high, based on the percentage of risk values obtained. A threshold value is set, so that the credit applicant below the threshold value is rejected and remaining credits are sanctioned. The sanctioned and rejected credit applicants are considered as ‘Good’ and ‘bad’ credits correspondingly.

REFERENCE

[1] A. Ghatge and P. Halkarnikar, "Ensemble Neural Network Strategy for Predicting Credit Default Evaluation."

[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. M. Mazid, S. Ali, and K. Tickle, "Improved C4. 5 algorithm for rule based classification," in Proceedings of the 9th WSEAS international conference on Artificial intelligence, knowledge engineering and data bases, 2010, pp. 296-301.

[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] S. Pourdarab, A. Nadali, and H. E. Nosratabadi, "A Hybrid Method for Credit Risk Assessment of Bank Customers."

[17] Ms. Neethu Baby and Mrs. Priyanka, "Customer Classification And prediction Based on Data Mining Technique", international Journal of Emerging Technology and Advanced Engineering, Research vol. 2, 2012.