# ERPCA: A Novel Approach for Risk Evaluation of Multidimensional Risk Prediction Clustering Algorithm

K. Kala
Research Scholar, Manonmaniam Sundaranar University, Tirunelveli
E-mail: kasinathkala1971@yahoo.co.in

Dr. E. Ramaraj
Professor, Department of Computer Science and Engineering
Alagappa University, Karaikudi
E-mail: eramaraj@rediffmail.com

***Abstract*—Clustering is a data mining technique used to place data elements into related groups without advance knowledge of the group definitions. In this paper clustering is employed to support efficient decision making by clustering mass storage data available in banks. Risk assessment is important task of Banks, as the failure and success of the Bank depends largely on banks' ability to evaluate the risk properly. The key problem consists of distinguishing, salubrious (good) and delinquent (bad) customers. Using a novel approach is proposed for risk evaluation of multidimensional risk prediction clustering algorithm (ERPCA). A risk evaluation process is used to determine the good and bad loan applicants. Feature subsets extraction is laid down on the multidimensional data using Information gain technique, to select the valuable attributes. Rules formation is done for each type of loans to avoid redundancy. In order to increase the accuracy of risk computation the risk assessment is performed in two levels, primary and secondary. This method allows for finding percentage of risk to determine whether loan can be sanctioned to a customer or not. This paper mainly concentrates on the clustering of the multidimensional data for efficient risk prediction.**

***Index Terms—Data mining, feature subset selection, information gain, clustering 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 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 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.

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 (ERPCA) 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 evaluation is an important and interesting management problem in financial analysis. *Francesca et al* proposed a time hazard model for a population of loans that involve different probability of default considering conjointly the explanatory variables and the time when the default occurs. Considering jointly the time and the risk factors a probability of default has been modeled for two main groups of loans: 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 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 was proposed by *Zakrzewska et al.* 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. *Bhasin et al* proposed a paper 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 concepts and tools of data mining are analyzed in this. Rule interestingness measures are discussed and a new 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 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. *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. *Kabir et al* 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. According to *Bodla et al* 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 was proposed by *Raghavan et al.*

The data warehousing solution should effectively interface with the transaction system like risks systems and core banking lists to collate data. *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 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. An Individual Credit Risk Evaluation System (ICRES) using data mining technology was proposed by *Liu et al.* The information gain method is used to screen the alternative indicators that have greater impact on the classification prediction. *Azhagusundari et al* 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. *Lopez et al* proposed an classification via clustering approach to predict the student's final marks in a university course on the basis of forum data. The motive of this paper is to predict the students pass/fail in the course using the classification via clustering approach. A meta-classifier applied for classification, uses a cluster for classification approach based on the assumption that each cluster corresponds to a class. The accuracy of using this approach is then compared with traditional classification algorithm. *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 non-modifiable attributes includes age, sex and family history of the premature. The modifiable factors include smoking before the event, history of hypertension and history of diabetes. 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. This 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 and consumes less time for execution. 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.

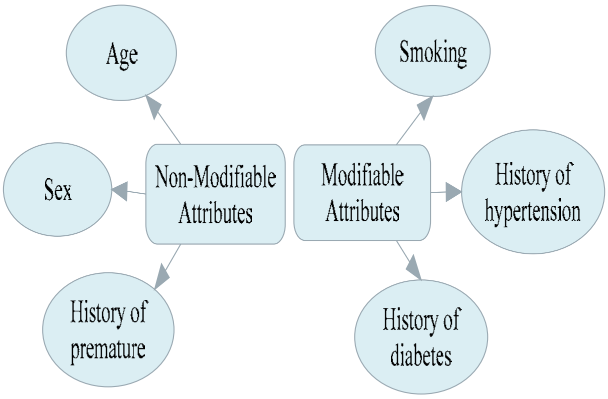


Fig 1: Attributes used in the existing method

## 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 (ERPCA) 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 (2). In this paper, the details (attributes) of the customers applying loan are collected and feature extraction is made on these attributes using info gain. Rules are formed for each loan type like (personal loan, bike loan, car loan, house loan, business loan). Risk assessment is done in two levels and finally the loan applicants are clustered based on the prediction as good or bad loan appliers.

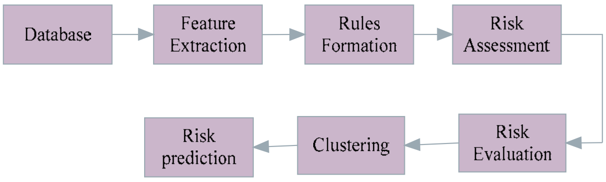


Figure 2: Overall Flow of the proposed work

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 in such a manner suitable for data mining. The dataset contains attributes like Age, Sector, Years of Experience, Property, marital status, Nominee, 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.

### A. *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.

### B. *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 the outcome of an information source is not known. Entropy is the measurement of how much information not known. 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 investigate

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

Where

$$p_i = probability$$

t = Class

D =Dataset

m =number of class values

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

|D| = total number of observations in dataset D

$D_j$ = 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.

### C. *Rules Formation*

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, 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. Rule list are framed by the bank as shown in the table 1.

### D. *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, percentage of risk is calculated. Percentage of risk is calculated using the equation (4). In this, greater weight age is given to Level I risk and lower weight age to level II risk.

Percentage of risk = (1- risk)*100 (4)

Where

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

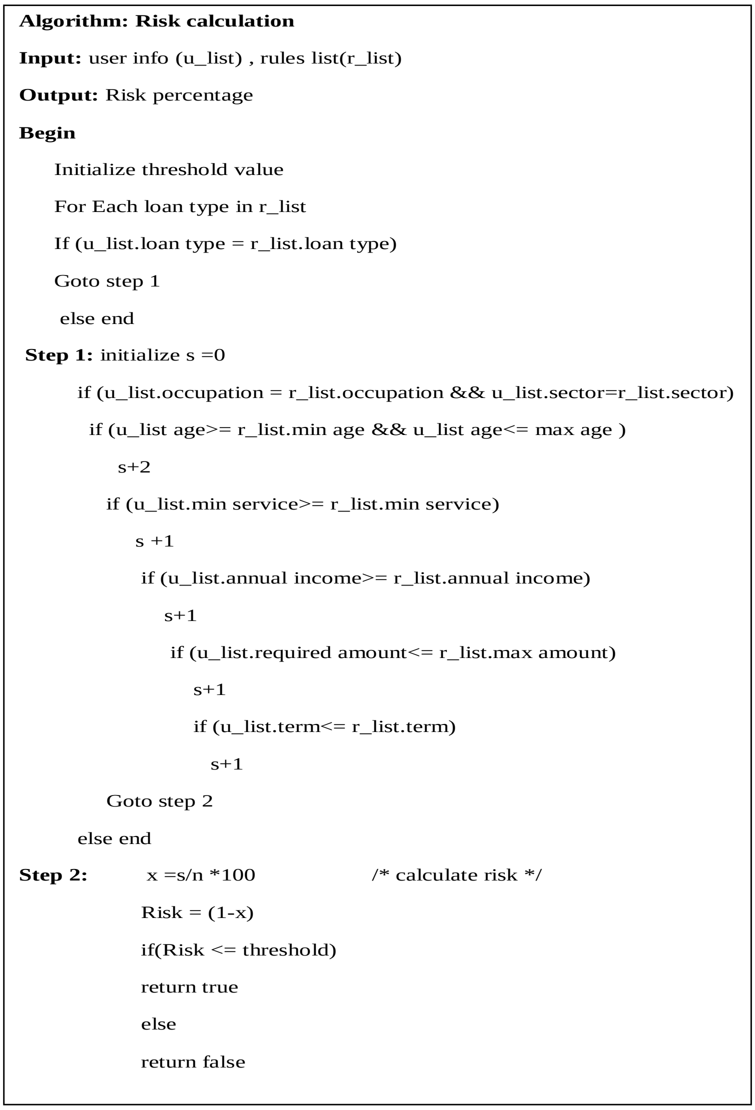
Percentage of risk is calculated for all the customers applied for loan, to evaluate whether loan can be sanctioned to a particular customer or not. Then the customers are classified, based on the percentage of risk obtained.

According to the algorithm 1, 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.

### *E. Clustering*

Clustering groups the set of objects and finds whether there is some relationship between the objects. Associative clustering algorithm (ERPCA) used in this work, effectively and efficiently mine clusters from massive and high dimensional numerical databases.

In this clustering, group of data elements can belong to more than one cluster, which is associated with each element is a set of membership levels. It indicates the strength of the association between that data element and a particular cluster. Clustering focus on assigning these membership levels and then using those membership levels to assign data elements to one or more clusters. Using ERPCA algorithm, three vectors can be taken into consideration. For e.g. percentage of risk, values can be categorized as low, medium and high vectors as depicted in clustering algorithm. Based on these vectors, the risk assessed data is then clustered. From the algorithm 2, it is seen that the classified multidimensional data are clustered by implementing proposed clustering. 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.



```
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
```

Algorithm 1: Rules prediction and Risk evaluation

Table 1: Rules Predicted from database

| 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 |

```
Input: Cluster (x),  ∈, L1,L2,M1,M2,H   /* x= % of risk values, ∈→ threshold */

Begin
 Clusters $w_k(x)$ = coefficients
Repeat until when $w_k(x) < \in$ ,
 Center for each x,
          $C_k = \frac{\sum_x w_k(x)x}{\sum_x w_k(x)}$                              (1)
 For each x
    Coefficient,
          $w_k(x) = \frac{1}{\sum_j (\frac{d(center\ k,x)}{d(center\ j,x)})^2 /(m-1)}$ ,   m =2        (2)
      Sub equation (2) in (1)
           If ($C_k \geq L_1$ && $C_k \leq L_2$) then
    $C_k = Cl_{d1}$
         Else if ($C_k \geq M_1$ && $C_k \leq M_2$) then
    $C_k = Cl_{d2}$
         Else if ($C_k > H$) then
    $C_k = Cl_{d3}$
       End if
   End For
 Collect all the clusters $Cl_{d1}, Cl_{d2}, Cl_{d3}$
   End
```

Algorithm 2: Proposed ERPCA algorithm

### F. *Risk prediction*

For the loan sanction, a threshold value of 35 percentage of risk is set. Risk is predicted based on this value, i.e. a lender can decide whether to sanction loan or not i.e. if the percentage of risk for a customer is greater than 35 percentage, 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.

## 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 seeked 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, the entropy for the feature selection is estimated. 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 3.depicts the number of features present before and after feature extraction process.

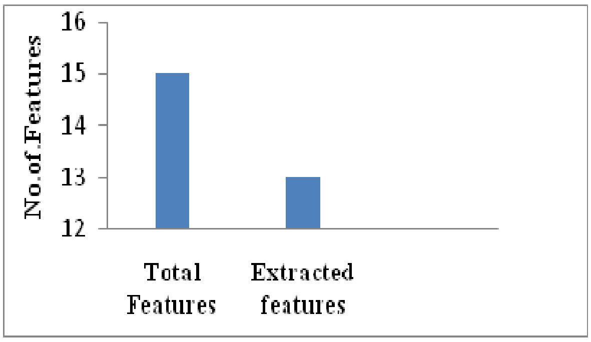


Figure 3: Features present before and after Extraction process

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 percentage of risk is calculated. Threshold value is fixed, so that the customers with percentage 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 percentage of risk value calculated, customers are classified as Low, medium and high as shown in figure 4.

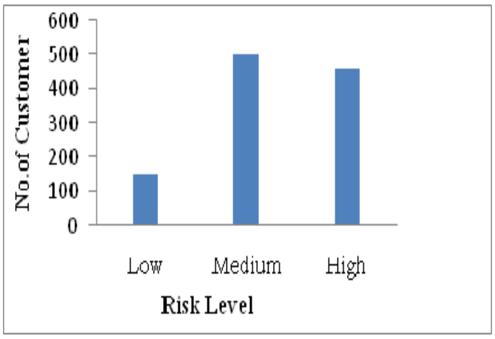


Figure 4: Risk levels obtained after classification

Finally, based on the percentage of risk value, loan applicants are separated as good credits and bad credits as shown in the figure 5.

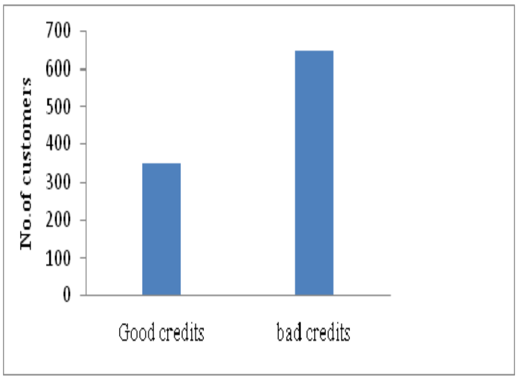


Figure 5: Illustration of good and bad credits

*Comparison with the existing system*

The proposed model (ERPCA) is compared with an existing technique Coronary heart disease risk evaluation (Risk evaluation-CHD) was proposed by *Anbarasi et al*. 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 (ERPCA) framework evaluates the risk in the given set of documents with greater accuracy and consumes less time than the existing technique.

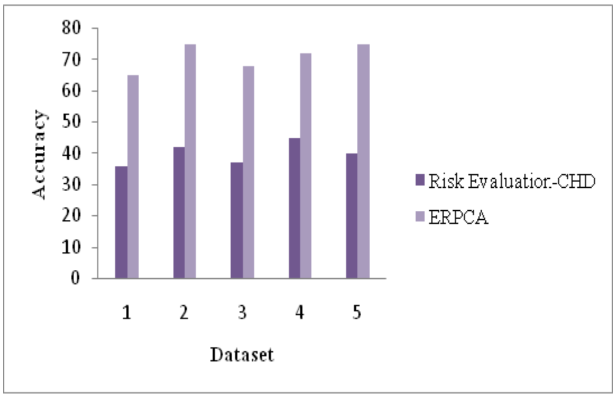


Figure 6: Accuracy comparison for Existing vs. proposed

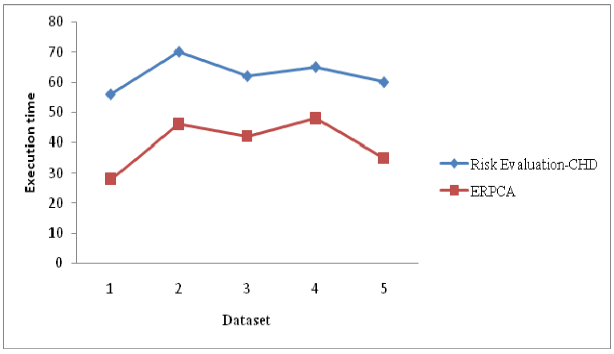


Figure 7: Execution time comparison for Existing vs. Proposed Method

## V. CONCLUSION

Risk Assessment is the crucial task in the Banking industry**.** This paper proposes a framework (ERPCA) 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. Clustering 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. Experimental results have shown that proposed method predicts the appropriate accuracy and also reveals that it consumes less execution time than the existing method.

Though the results obtained here are encouraging, future exploration is still necessary. With respect to future enhancements, this work is open to discover the cubes grade problem, which has implication in large multidimensional.

## 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.