CREDIT RISK CALCULATION AND LOAN REPAYMENT SUGGESTION FOR BANK DATA USING DATA MINING

Dr.V.KANNAN
Athma Educational Institution, Coimbatore, Tamil Nadu, India
drkannanv.india@gmail.com

Abstract: Data mining plays a vital role in almost all types of domain and application as of now the banking domain proves to be the upcoming and challenging domain. As banking have multi-dimensional data and mass data size these challenges occurs. The customer size is drastically increasing and so the credit card requests and the loan approval became a tough process to handle. Inorder to handle bank customer profile efficiently, the decisions on credit card approval and to provide the prediction of risk on loan repayment the proposed system establishes a new data mining technique. Eventhough several solutions are provided by the existing system for handling the bank dataset and to predict the credit card risks, the system will suffer when the data size is huge and the attributes are dynamic. Thus the proposed system introduces and creates a new decision support system and some data classifier to handle such risk hidden data. Two data mining techniques were proposed in this paper to support loan decisions for the banks. Loan applications from different customers are collected and are evaluated to check whether the applicant will pay his/her loan or not. Further analysis and decision support will be performed on the selected applications. The proposed credit risk calculation and decision support system is named as (Ek-EDT) extended k-means and enhanced decision tree for fast decision making and data management in banking sector. This helps to find the effective features and classifies the applicants into risky and normal category with effective decisions and suggestions.

Keywords: banking, risk assignment, data mining, decision support system.

1. INTRODUCTION

Data mining techniques are emerging as the most widespread domain nowadays. This is because of the extensive accessibility of huge quantity of data and the need for transforming those data into knowledge. In the current financial sector, the core banking model and cut throat competition is making the banks to struggle inorder to gain a competitive edge over each other. In this modern banking world, the face to face interaction with customer doesn't exists. Huge amounts of data is been collected by the banking systems on day to day basis like customer information, their transaction details i.e., deposits and withdrawals, loans, risk profiles, credit card details, credit limit and collateral details related information. On daily basis, the bank refers thousands of decisions to support the customers. The ability to generate, capture and store data of the customers has been tremendously increased in recent few years. The data collected are with precious information's. And the wide availability of huge amounts of data and the need for transforming such data into knowledge encourages IT industry to use data mining.

The major business process in banks is lending. And the most critical and necessary factor in the banking world is Trust Risk Management. The credit risk management is to be maintained properly inorder to avoid huge losses for the banks and this prevents the lending process very tough for the banks. Hence, Data mining techniques are greatly used in the banking industry which helps them to compete in the market and to provide the right product to the right customer with less risk factor. Credit risks can lead to the risk of loss and the loan defaults are the major source of risk encountered by banking industry. Classification and prediction are the data mining techniques that can be applied to overcome these risk factors to a great extent.

An effective prediction model is been introduced in this paper for the bankers. It facilitates them in the prediction of the credible customers who have applied for loan. Decision Tree Induction implied the Data Mining Algorithm concept is applied to predict the attributes relevant for credibility. A prototype of the model is described in this thesis. It is used by the organizations in making the right decision to approve or reject the loan request of the customers and also to segment them into golden, silver, and risky customer

types. The following chapters discuss about the introduction of data mining and other concepts in details.

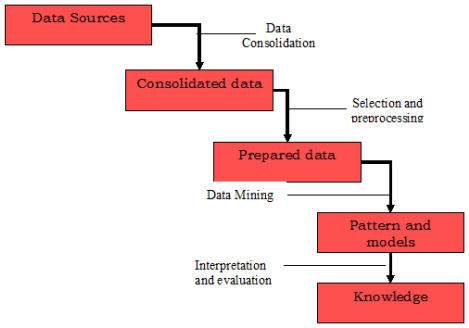


Figure1: Architecture of Data Mining

2. CLUSTERING CONCEPTS

Classifying the objects into different groups i.e., the partitioning of a data set into subsets (clusters) using data mining techniques is performed so that the data in each subset (ideally) share some common trait - often state according to some defined distance measure. The statistical data analysis, which is used in many fields, including machine learning, pattern recognition, image analysis, bioinformatics and data mining is the common technique of Data clustering. The concepts of data clustering or clustering are homogenous with the concepts of cluster analysis, automatic classification, numerical taxonomy, Botryology and typological analysis.

2.1. Introduction

The process of partitioning the set of data or objects in a set of meaningful sub-classes, called Clusters is the technique involved in the Clustering process. Without advanced knowledge of the group definitions, the data elements are placed into related groups by this clustering process which is a data mining technique. Cluster of objects are created with somehow similarly characterized objects by involving the Data clustering method. The definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". Thus a Cluster is a collection of objects which behaves "similar" between them and is "dissimilar" to the objects belonging to other clusters. The users can show this with a simple graphical example:

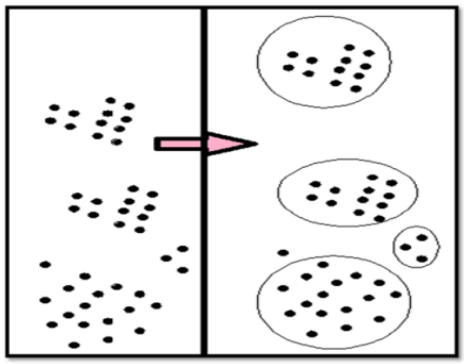


Figure 2: Clustering process

In the above case the identification of 4 clusters is made performed into which the data can be divided; here the similarity criterion is distance: two or more objects belong to the same cluster if they are "close" according to a given distance (in this case geometrical distance) and this is termed as distance-based clustering. Another kind of clustering is called conceptual clustering, in this two or more objects belong to the same cluster if this one defines a concept common to all that objects. In other words, objects are grouped when they fit to the descriptive concepts, but not only depending on the simple similarity measures.

2.1.1 Types of Clustering Methods:

Among the availability of multiple clustering methods, each of their methods develops different grouping of a dataset. So the choice of a particular clustering method will depend on the type of output desired, the known performance of method with particular types of data, the hardware and software facilities available and the size of the dataset. Based on the cluster structure they establish, the clustering methods can be classified into two categories as non-hierarchical and hierarchical method. First in the non-hierarchical methods a dataset is divided into N number of objects that in turn makes M clusters, with or without overlap. These non-hierarchical methods sometimes it can be divided further into partitioning methods, in which the classes are mutually exclusive and in the less common clumping method may also occur in which overlap is allowed. Each object is a member of the cluster with

similar characteristics; however the threshold of similarity has to be defined. Secondly, the hierarchical methods which develop a set of nested clusters. These nested clusters have multiple pair of objects or clusters in which each pair is progressively nested in to a larger cluster until only one cluster remains. And these hierarchical methods can be further sub-divided into agglomerative or divisive methods. In agglomerative methods, the hierarchy can be built in a series of N-1 agglomerations, or by Fusion which is performed by pairs of objects that begin with the un-clustered dataset. Another method called less common divisive method, which begins with all objects in to a single cluster and at each N-1 step it divide some clusters into two smaller clusters, until each object resides in its own cluster. Certain important Data Clustering Methods are described below.

2.1.2 Partitioning Methods

The partitioning methods generally result in developing a set of M clusters in which each object belongs to one cluster. Each cluster may be represented with some sort of summary description about all the objects contained in a cluster. And this way of representation is done either with centroid or by cluster representation. And this way of representation of cluster depends on the type of the object which is being clustered. Cluster representation is done if real-valued data is available by calculating the arithmetic mean of the attribute vectors for all objects within a cluster. Alternatively different kinds of centroid are implemented in other cases like representation of cluster of documents by a list of those keywords that occur in some minimum number of documents within a cluster. If the number of the clusters is large, then again the centroids could be further clustered to produce a hierarchy within a dataset.

Single Pass:

A very simple partition method, used to create a partitioned dataset is the single pass method which is as follows:

1. Make the first object which is described in the centroid for the first cluster.
2. For the next object, calculate the similarity S using each existing cluster centroid along with some of the similarity coefficient.
3. If the calculated S value is greater than some specified threshold value, then add the objectto the corresponding cluster and recalculate the centroid else use the object to initiate a new cluster. If any objects remain to be clustered, return to step 2.

As its name itself implies that this method requires only one pass through the dataset and so the time requirements are typically of order O (N log N) for order O (log N) clusters. This makes it this clustering method very efficient in the serial processor. The major drawback of this method is that the resulting clusters are not independent of the order, in which the documents are processed. As the first clusters formed are usually larger than those created later in the clustering run.

2.1.3 Hierarchical Agglomerative methods

The hierarchical agglomerative method of clustering is wide commonly used. The general algorithm is used in the construction of a hierarchical agglomerative method by the following steps involved.

1. Merging of 2 closest objects into a single cluster by finding them.
2. Finding and merging the next two closest points, where this point could be either an individual object or a cluster of objects.
3. If more number of clusters remains, then return to step 2.

Depending on the definition used for the identification of the closest pair of points individual methods are characterized which in turn helps to describe the new cluster when two clusters are merged. There are few general approaches to implement this algorithm, those are second matrix approach and stored data approach which are discussed below:

In the second matrix approach, an N*N matrix that contains all pairwise distance values are created first, and updated as the new clusters are formed. This approach has at least an O(n*n) time requirement. The requirement can be raised to $O(n^3)$ even when the dissimilarity matrix is serial scan simply then these are used to identify the points that are needed to be fused in each agglomeration, where a serious limitation for large N occurs.

The stored data approach requires the recalculation of pairwise dissimilarity values for each of the N-1 agglomerations, and the O (N) space requirement could

be achieved at the expense of an $O(N^3)$ time requirement.

2.2 Use of Clustering in Data Mining

Preferably the initial step in data mining analysis is the process of clustering. As this process identifies the group of related records that can be used as an initiating point for exploring further relationships. The population segmentation models is been supported by this technique for their development, like by using the demographic-based customer segmentation. Using the standard analytical methods and other data mining techniques additional analysis can be performed. In addition, the characteristics of these segments can also be determined with respect to some desired outcome. Let us consider an example of buying habits of multiple population segments, this data can be used to determine which segments to target for a new sales campaign. Another process like a company's analyzing process, in which the scope of sale in their wide variety of products is determined so that the product with better scope and which is lacking can classified. These processes of analyzing can be enhanced by data mining techniques. This system when clusters the products of less availability of sales then such products could be checked alone rather than comparing all the sales value of the products. And this actually facilitates the mining process.

3. CLASSIFICATION

Classification is a data mining (machine learning) technique used in the data instances inorder to predict group membership. Classification can also perform the data mining function that assigns the target categories or classes with the items collected. The goal of this classification technique is to accurately predict the target class for each case in the data.

For example, a classification model could be used in classifying the loan applicants into three categories such as low, medium or high Trust risks. Classification models can be tested by comparing the known target values to the predicted values in a set of test data.

The classification project based on the historical data is typically divided into two data sets: one for building the model and other for testing the model. The data mining classification algorithm and machine learning when provided with a set of inputs, they come up with a specific class associated with those inputs.

Classification is the problem identifier for the set of categories i.e., sub-populations where a new observation belongs. This identification process functions based on the training set of data containing observations (or instances) whose category membership is known. Classification method that functions as an identifier is classifier which depends greatly on the characteristics of the data to be classified.

It's impossible to solve all the problems by implementing a single classifier technique (a phenomenon that may be explained by the no-free-lunch theorem). The classifier performance can be verified by performing multiple empirical tests and so to find the characteristics of data that determine classifier performance. The processes of determining a suitable classifier for a given problem is still more an art than a science.

3.1 Decision tree induction

The process of learning of the decision trees from class-labeled training Tuples is called as Decision tree induction. This in induction process also helps in learning of the decision trees from class-labeled training tuples. The Decision tree induction has simple and fast learning and classification steps. And it also ensures in providing better accuracy. But the success rate of this method depends on the type of data present in hand.

The field of medicine, manufacturing and production, financial analysis, astronomy, and molecular biology are the areas of application of Data tree induction where classification technique is performed. Several commercial rule induction systems obtainted their basic features using Decision trees.

A decision tree is a flowchart-like tree structure with topmost in the tree as root node in which each leaf node (or terminal node) holds a class label, each branch represents an outcome of the test and each internal node (non-leaf node) denotes a test on an attribute. For the question "How is decision trees used for classification?" The answer will be- for a given a tuple, X, for which the associated class label is unknown, the attribute values of the tuple are tested against the decision tree.

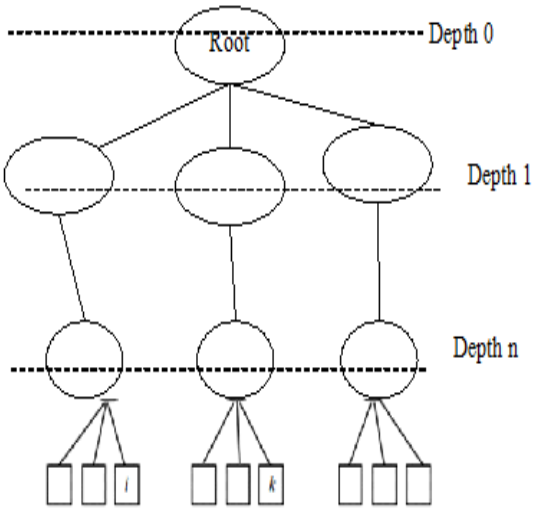


Figure 3: decision tree model

By recommending only for the customers who are likely to buy recommended products, the recommendation errors can be minimized using the Decision tree induction. The false positives of the poor recommendation could be avoided by the solution of recommending only for the customers who are likely to buy recommended products. Such customers were selected in this phase based on the decision tree induction. The model set and the score set are generated from the customer records, that could be used in the decision tree induction.

3.1.1 Advantages of using decision learning tree algorithms are:

- The unobserved instances could be generalized in a better way by examining the attribute value pair single time in the training data.
- They perform computation in an effective way and that efficiency is proportional to the number of training instances observed.
- The tree interpretation method detail view of understanding how to classify instances based on attributes that are arranged on the basis of information they provide and also makes the classification process a self-evident.
- The operation of decision tree is based on the algorithm called ID3 or C4.5. Based on the information (information gain) obtained from the training instances were used to build a tree and the same data is used to classify the test data. ID3 algorithm generally uses nominal attributes for classification with no missing values. ID3 can even work well on datasets with missing attribute values to certain extent.

- The continuous and discrete attributes are handled by C4.5. While handling the data and it allows the missing attribute values to be marked as (?). The missing attribute values are not used simply in gain and entropy calculations. Decision tree are self-explanatory. And so it could be easily converted to a set of rules, which could be used in credit evaluation process. Among data mining tools, the artificial neural networks are most commonly used.

3.1.2 Neural networks

For the tasks of classification, prediction, and clustering Neural networks were particularly used in the business applications. The three properties are utilized to characterize the neutral network: the Credit risks which points for the risk of loan defaults and loss and the computational property. These are the major source of risk encountered by banking industry.

Classification and prediction techniques of Data mining could be applied to overcome this to a great extent. This is an area which could show tremendous increase in the profit of the lender even with small-scale of performance improvement. This is because of the volume and quantity of the lending amounts. Nowadays, banks are recognizing the various advantages of data mining. Banks could perform the process of identifying the potentially useful information from the large amounts of data using the valuable tools of data mining. This can help banks to gain a clear advantage over its competitors. Thus Data mining can help banks in better understanding of the vast volume of data collected by the customer management systems.

3.2 Motivation

The financial institutions and customers are growing in huge numbers at present situation. This leads to the development of multiple applications and manpower for each and every financial institution. And such banks and institutions are providing different loan options and financial services to their customers by performing the background verification and approval process. Before sanctioning the loan amount to the applicant, there is a need to analyze their profile. From the profile, the bank ensures an idea about the applicants whether they can

repay the loan amount or not. And this process of verification is performed in the bank manually. Here comes the problem and so this manual work should be reduced and necessary research work should be performed in the banking sector. This could be performed by the data mining and decision making algorithms that should be created that could solve the issues related to the data management.

The dimensionality problem is the key challenge in this banking domain, which denotes the presence of huge volume of data. Data mining techniques are computationally efficient to handle the large sized inputs obtainted from the banking domain. In this proposed system, the concept of classification, prediction system, interpretation and Effective analysis is been developed effectively. Already existed ata mining approaches have not combined extended K-means algorithms and decision making systems.

- Higher level of preprocessing is required in the real time in banking systems dataset.
- Classification and Decision Tree should be balanced for the purpose of integration.
- A generalized Data mining framework can't be used for all the different systems available.
- For the quick decision making, banking domains need the support of the data mining techniques.

3.3 Aim and objective

The main aim of the proposed system is to discover an effective algorithm that can handle huge set of attributes and values in the financial data. The main objective of the proposed system is to utilize the popular and effective detection method strategies inorder to find the fast and minimum iterative customer classification and loan risk prediction. The risk prediction and the risky customer detection techniques face another problem with an increase in uncertainty results and difficulty in the identification of the exact value of the risk. So this work focused mainly in providing a major contribution towards the design of automatic methods for the discovery of properties which reduces the iteration in the decision making problem.

To predict unknown or future values in the banking dataset the quality and prediction assessment is included as the objective of data mining technique. The quality assessment refers to assessing the quality of values predicted. Predictive modeling is used widely in deriving the analytical information for assessing the banking dataset to provide information to the banking team with the likelihood that a patient or a user specific action. The actions predicted by the predictive modeling comprises of how the diseases are classified and how effective the diseases of the patients are predicted, how the disease are classified.

4. DECISION TREE ALGORITHMS

Decision trees are the tree types that classify the instance using the feature values inorder to sort them. Each branch of the decision tree represents a value that the node can assume whereas each node in the decision tree represents a feature in an instance that has to be classified. The classification of the instance initiates at the root node and they are sorted depending on their feature values. An example of a decision tree for the training set of Table I. For a certain period of time, multiple decision tree algorithms are developed by the researchers. Their work included the enhancement in performance and ability to handle various types of data. Few important algorithms are discussed below.

CHID: The fundamental decision tree learning algorithm called CHi-squared Automatic Interaction Detector or CHAID was developed by Gordon V Kass [4] in 1980. It has the features like easy to interpret, easy to handle and can be used for classification and detection of interaction between variables. The AID (Automatic Interaction Detector) and THAID (Theta Automatic Interaction Detector) procedures is the extended as CHAID. The working principle of CHAID is the adjusted significance testing. The best attribute for splitting the node is selected after detecting the interaction between the variables thus creating a child node which contains the collection of homogeneous values of the selected attribute. This method does not suggest any pruning method but it can handle the missing values

CART: Breiman et al. [5] proposed the Classification and regression tree or simply CART which constructs binary trees that can also be referred to as the Hierarchical Optimal Discriminate Analysis (HODA). Depending on whether the dependent variable is categorical or numeric, the classification or regression trees are developed respectively and this makes CART

a non-parametric decision tree learning technique. As the word binary implies that a node in a decision tree can only be splitted into two groups. CART performs the selection of attributes by using the gini index as the impurity measurement. The node record is splitted with the attribute in which the largest reduction in impurity is identified. CART handles missing attribute values and also accepts data with numerical or categorical values. It uses cost-complexity pruning method and also generate the regression trees.

- **ID3:** Iterative Dichotomiser 3 or ID3 decision tree algorithm was developed by Quinlan [6]. In the decision tree method, the suitable property for each node of a generated decision tree is determined using the information gain approach. Thus, test attribute of the current node is been selected of attribute with the highest information gain i.e., entropy reduction in the level of maximum. In this way, the the training sample subset is classified with the information obtained from later on partitioning will be the smallest. It shows that the use of this property to partition the sample set contained in current node will make the mixture degree of different types for all generated sample subsets reduce to a minimum. Therefore, the use of such an information theory approach i.e., ID3 will effectively reduce the required dividing number of object classification.

C4.5: A decision tree developed by Ross Quinlan is based on the algorithm of C4.5. This C4.5 is an extension of Quinlan's earlier ID3 algorithm. The C4.5 is often referred as a statistical classifier [7] this is because the decision trees generated by C4.5 algorithm could be used for classification

- The splitting criteria used by C4.5 algorithm is the information gain. The categorical or numerical values can be given as the input value. Threshold values are generated to handle the continuous values. Generated threshold values helps to divide the attributes with values above the threshold and values equal to or below the threshold. C4.5algorithm can easily handle missing values and the missing attribute values are not utilized in gain calculations by C4.5.

- **C5.0/Sec 5:** The extension of C4.5 algorithm and ID3 is termed as C5.0 algorithm. It is the classification algorithm which can be applied in the big data set and also better than C4.5 on the speed, memory and the efficiency. The maximum information gain is provided by C5.0 model that functions by splitting the samples based on the field. The C5.0 model can also split samples on the basis of the biggest information gain field. The sample subset that is obtainted from the former split will be splitted afterwards. The process will continue until the sample subset cannot be split further and is usually according to another field. Finally the examination of the lowest level split is performed. Those sample subsets which doesn't have remarkable contribution to the model will be rejected. C5.0 can be easily handled for the multi value attribute and missing attribute from data set [8].

- **Hunt's Algorithm:** Hunt's algorithm generates a Decision tree by divides and conquers approach or by top-down approach. The sample otherwise termed as row data contains more than one class that use an attribute test inorder to split the data into smaller subsets. Hunt's algorithm establishes an optimal split for every stage according to some threshold value in a greedy fashion [9].The proposed system implies the decision tree based algorithm in the following domain.

The research and practice in the field of Medicine are the important areas of application for decision tree techniques. Decision tree is the most useful technique in finding various decisions on banking. It is also used in the loan repayment probability detection.

Table 1 Comparisons between different Decision Tree Algorithm.

4.1 EXISTING C4.5 ALGORITHM

Let the classes be denoted as $\{F_1, F_2,\ldots,F_k\}$. There are three possibilities available for the content in the set of training samples T in the given node of decision tree:

- T contains one or more samples and all these belong to a single class F_j. The decision tree for K is a leaf identifying class F_j.
- **C4.5:** A well-known algorithm used to generate a decision tree is C4.5. It is an extension of the ID3

algorithm that likely helps in overcoming its disadvantages. C4.5 algorithm which generates the decision tree is also referred to as a statistical classifier this is because this algorithm can be used in the classification technique.

4.1.1 Disadvantages of C4.5

C4.5 algorithm constructs empty branches and it is the most crucial step for rule generation in C4.5.We had found many nodes with zero values or close to zero values. And these values can neither contribute to generate rules nor help to construct any class for classification task.

Rather it makes the tree bigger and more complex, over fitting happens when algorithm model picks up data with uncommon characteristics.

5. METHODOLOGY

This chapter describes about the detailed process of Ek and EDT algorithms. And this study is limited to the specific banking sector. The process involved in this study is as follows: Initially the system collects numerous customer records from the UCI repository. This data set includes several fields which are related to the Coimbatore district. The details are enclosed in the chapter 1. The data set obtainted may also contain some missing values. So initially the data preprocessing is performed by the system. Finally, the pre-processed data will be transformed into a suitable format so that data mining techniques can be applied. The following fig 3.1shows the overall process involved in the proposed system. The initial stage of the proposed work is the process of initialization, where the data collection, selection and transformation process is established.

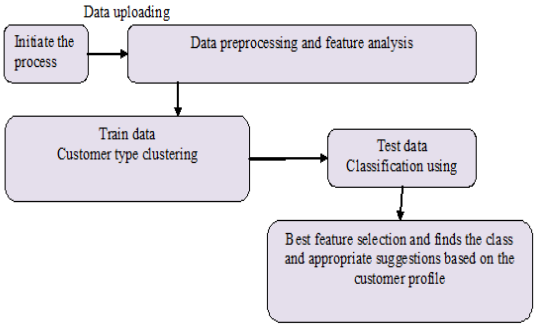


Figure 4: Overall processes of Ek-EDT

5.1 DATA PREPROCESSING

Data preprocessing steps are applied on the new set of data like the customer loan application and they are converted to categorical values by applying filters and by using unsupervised clustering algorithm named as an enhanced extended K-means. After the operations are carried out, the total number of input instances of the individual locations is presented for analysis.

The attributes in the bank data set are filtered and the selection of relevant attributes needed for prediction is performed. After this process, the incomplete and noisy records in the dataset are removed and prepared for the mining process.

5.2 DATA SELECTION

At this stage of data selection, relevant data for the analysis will be decided and are retrieved from the extracted customer data set. The extracted customer data set had several attributes like their type and description .

Input variables:

bank client data:

1 - Age (numeric)

2-job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management' ,'retired','self-employed','services','student','technician','unemployed','unknown')

3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)

4 - education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')

5 - default: has credit in default? (categorical: 'no','yes','unknown')

6 - housing: has housing loan? (categorical: 'no','yes','unknown')

7 - loan: has personal loan? (categorical: 'no','yes','unknown')

related with the last contact of the current campaign:

8 - contact: contact communication type (categorical: 'cellular','telephone')

9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')

10 - day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')

11 - duration: last contact duration, in seconds (numeric). Importantly this attribute highly affects the output target this is because if duration=0 then y='no'. Yet, the duration is not known before a call is performed. Also, after the end of each call y is obviously known. Thus, duration input should only be included only for benchmark purposes and should be neglected if the intention is to have a realistic predictive model.

other attributes:

12 - campaign: number of contacts performed during this campaign and for the respective client (numeric, includes last contact)

13 - pdays: number of days passed after the client was last contacted from the previous campaign (numeric; 999 means client was not previously contacted)

14 - previous: number of contacts performed before this campaign and for the respective client (numeric)

15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure', 'nonexistent', 'success')

social and economic context attributes

16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)

17 - cons.price.idx: consumer price index - monthly indicator (numeric)

18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)

19 - euribor3m: euribor 3 month rate - daily indicator (numeric)

20 - nr.employed: number of employees - quarterly indicator (numeric)

21-gender
22-income
23-spouse income
24-number of children

5.2.1 Customer data descriptions

In this stage the system is developed in an efficient and user-friendly manner. And so it can also support the users with less technical knowledge to carry out the process comfortably. The system provides the most relevant attributes that helps in determining whether to approve or reject the loan application and that aids in predicting the credibility of future customers.

The data sets are classified into testing data sets and training data sets that are used in this research work. All the training dataset by default considers only the banks' guidelines for accounting the personal credit approval. For this process 1000 customers data are retrieved. The data required for the current study was collected from the UCI repository. It consists of one dependent variable and multiple independent variables. Variables are the conditions or the characteristics that the investigator can manipulate, controls and observe thus it is necessary to optimize the variables by using Ek-EDT. Variables can be classified as dependent and independent variables. An independent variable is the condition or the characteristic that affects one or more dependent variables by its size, number, length or whichever attribute that exists independently and it is also not affected by the other variable. A dependent variable is the variable that changes as a result of changes in the independent variable. Independent Variables data set can be used to build a model, which consists of a decision tree model EDT inorder to predict whether a applicant's can pay their credit is paid or unpaid. This chapter can use the decision tree node to classify observations by segmenting the data created according to a series of simple rules and can use the entropy gain reduction method to build the tree. The regression node is incorporated to the data using the logistic regression model to the data. Thus, the EDT is successfully implemented to the banking domain.

5.2.2 Proposed algorithm using EK-EDT -

After the segmentation, the system performs the rule set definition by implementing the mean, median and variance concepts on the data. The Pearson distribution is been used to calculate the correlation in the data.

Algorithm EK -phase1 **(labeled example S, set of variables X)**

Input: A set S of labeled examples, a set X of variables.

Output: Feature set

1. Let B = empty.
2. Get the training data D into C subsets Dc by the class value c or customer profile attribute.
3. for each training data set Dc
   - Compute the Mean M (X_i;X_j) and the Mode (X_i;X_j) between each starting to end tuple of variables X_i and X_j.
   - Compute W (X_i) for each variable X_i.
   - For all variables X_i in X

- Add all the variables X_j with W (X_j) > W (X_i) to the parent set $_{Xi}$ of X_i.
- Add arcs from all the variables Xj in X_i to X_i.

- Add the resulting network B_c to B.

4. Return B.

In order to improvise the algorithms like C5 and C4.5, the Bank decision tree algorithm made that includes number of changes. Some of these are:

• The proposed system handles training data with missing values of attributes. So, the prediction will be more accurate and effective.

• Handling the changing values of data typed features.

• Based on the historical dataset, the Prediction probabilities are made.

• After the creation of decision tree pruning is performed.

• By modifying the process, the performance of the weak classifiers can be improved.

• Handling attributes with discrete and continuous values. Let the training data be a set R= R1, R2 ... of already classified samples. Each sample Di = Rl , R2... is a vector where Rl , R2 ... represent attributes or features of the sample.

The training data is a vector V = V1, V2..., where V1, V2... represents the class to which each sample belongs. At each node of the tree, C4.5 chooses one attribute of the data that most effectively splits data set of samples T into subsets that can be from one class or the other. As the result of choosing an attribute for splitting the data, the normalized information gain i.e., difference in entropy is retrieved. The attribute factor with the highest normalized information gain is considered to make the right decision for the selected test samples.

- **EK-EDT:** Various innovative features are developed by the system to establish new decision tree algorithm. The current study deals with the problem which is not handled in the literature i.e., the combinatorial optimization problem. So a new algorithm is created named as bank decision tree algorithm. The system performs an iterative method to improve the accuracy and efficiency and also aims at reducing the training complexity in the classification of Bank customer details.

An accurate EK-EDT training method is built by exploiting frequent items and rule set by EK-EDT. Consider a labeled structured data set D, a frequency, rule threshold, and a test case T. EK-EDT exploits the novel approach to estimate each class of D. EK-EDT is a novel iterative classifier. It evaluates and selects set of eligible items when a new test case T has to be classified. To perform efficient pattern retrieval, EK-EDT first performs data storage in a disk-based compact data representation. The followings are the major process included in this proposed study.

- Customer profile extraction
- Training phase
- Test Phase

In this section, the EK-EDT is thoroughly described and also represents the training phase of EK-EDT. The test phase of EK-EDT is based on the basic statistical evaluation process.

5.2.3 Phase 1: loan repayment prediction and risk analysis

The main process of the phase 1 is the optimal feature detection and the risk factor detection to the customer. When an attributes frequently occurs, that is the strongest attribute compared to other attributes. For customer classification, the system finds salary and other important features of the given loan data. The results of the phase1 are explained below.

5.2.4 Customer Classification using enhanced EK-EDT

Input: Customer features.

Output: Customer type and loan repayment risk

Steps:

1. Read training samples S.
2. For each features F in S do
   a. Calculate feature score fs=unique(FSi)
3. Order the fs desc and do
   a. For each class C in S, tfs=test(fs→FSi→Ci)
4. Return the scores
5. Find the max(tfs) and return the class Ck

Based on the feature scores and with the above algorithm, the system finds the appropriate class for the given features. For example the given input is (Customer profile: employment type, age, gender, spouse income, income, own house and car details, children count etc) then the output generated by the proposed system will be as follows. The detected class is unpayable.

Table 1: customer classification repayment classification score report

Customer id	Detected class	Score
1	Un payable	0.42709
2	Payable	0.4605
3	Un payable	0.902789

5.2.5 Phase 3: Bank loan risk based decision using the Customer profiles:

The large decision tree can be viewed as a set of rules which is easy to understand. EK-EDT algorithm gives acknowledgement on noise and missing data. This in turn solves the problem of over fitting and error pruning. In the classification technique, the EK-EDT classifier can anticipate by classifying the attributes which are relevant and not relevant.

This helps to find the best decision for different inputs based on location based information's. To allow efficient per-class pattern retrieval, the EK-EDT training phase builds a separate score level data to represent the training data to score compactly which belongs to each class. An effective classification technique is then exploited to efficiently retrieve the frequent cause and feature from the stored representation. The EK-EDT is an iterative data structure that is frequently used in the dataset. It is used to compactly represent the main memory transactional data sets. The initial frequency calculation process for EK-EDT algorithm has been used at the time of training. The Materialization Tree-based data representation allows EK-EDT to cope with large data sets.

Algorithm 1: EK-EDT Training Phase (D, C)

Input: the training set D

Output: EK-EDT Tree for each class belonging to the training class set C

1: for all ci in C do

2: rci = set of all training records belonging to class ci

3: calculate M=sum(Ci)/count(D_c)

4: end for

5: return rule r]

EK-EDT -Test Process:

Step 1: Generate the base conditions:

a) Get test data, training data and the rule set that supports the training phase of EK-EDT model
b) Get Threshold value(maximum value)
c) If the data satisfies the threshold then- add to the list
d) Else if the attribute performs the following
    a. Based on the value it captures the least possible values from the data
    b. Repeat until checking complete all features

Step 2: Combine the results of each attribute

Step 3: Find the maximum probability.

Step 4: Finally EK-EDT algorithm generates an optimal class of data using Extended K means approach by performing the test process.

The EK-EDT process also finds the impact of the suggested decision. This finds, whether the given suggestion increases the loan repayment for the location. The followings are the possible suggestion to increase the bank repayment.

6. RESULTS AND DISCUSSION

In this chapter, the efficiency of the algorithms is evaluated, in terms of time consumption against dimensionality d, number of traversal of tree, and tree generation threshold q under two distributions of different customers. This also evaluates the progressiveness of the methods under different datasets.

This section evaluates the proposed Credit risk analysis with EK-EDT data framework in terms of both tree traversal overhead and accuracy. We applied Credit risk analysis on sample customer's data for the experiments.

The above experiments represent the suggestions for the iterations based on the current customer type and details.

For the experiment, An Intel I3 2.2 GHz processor with 2 Gb RAM was used to measure the execution time and detection speed. Table 2 describes the execution time for varying dataset values and the accuracy for varying dataset value.

Table 2: execution Time comparison table

Algorithms	Time	100	300	500
J48		265	355	493
Credit risk analysis using EK-EDT		131	239	369

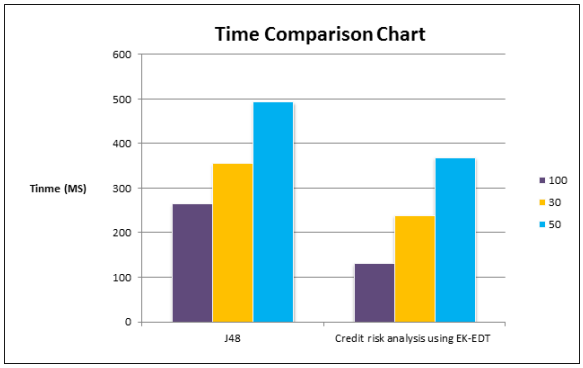


Figure 5: execution time comparison chart

The following figure 5 shows comparison between the existing J48 and Credit risk analysis EK-EDT. The experiment result shows that the proposed system is faster than the existing system.

This chapter compares the proposed EK-EDT with existing J48 and the C4.5 model, in terms of tree construction time.

7. CLASSIFIER ACCURACY MEASURES

The training data is used to derive a classifier or predictor. The accuracy prediction of the resulting learned model can lead to misleading and overoptimistic estimates due to overspecialization of the learning algorithm to the data. Instead, accuracy is better measured on a test set consisting of class-labeled tuples that were not used to train the model. The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified classifier. In the pattern recognition literature, it's also referred to as the overall recognition rate of the classifier which reflects how well the classifier recognizes tuples of various classes.

The error rate or misclassification rate of a classifier, M, which is simply termed as 1-Acc(M), where Acc(M) is the accuracy of M. If we want to use the training set to estimate the error rate of model, this quantity is known as the resubstitution error. This error estimation is optimistic of the true error rate because the model is not tested on any samples that are unknown to it. The confusion matrix is a useful tool for analyzing how well our classifier can recognize tuples of different classes. For a given m classes, a confusion matrix is a table of at least size m by m. An entry, $CM_{i,j}$ in the first m rows and m columns indicates the number of tuples of class that are labeled by the classifier as class j. For a classifier to have good accuracy, most of the tuples would be represented along the diagonal of the confusion matrix, from entry $CM_{1,1}$ to entry $CM_{m,m}$, with the rest of the entries being close to zero. The table may have additional rows or columns to provide totals or for recognition rates per class.

Given two classes, in terms of positive tuples like credit risk_occurrence = yes versus negative tuples like credit risk_occurrence = no. True positives refer to the positive tuples that are correctly labeled by the classifier, while true negatives are the negative tuples that are correctly labeled by the classifier. False positives are the negative tuples that are incorrectly labeled (tuples of class credit risk_occurrence = no for which the classifier predicted credit risk_occurrence = yes). Similarly, false negatives are the positive tuples that are incorrectly labeled (tuples of class credit risk_occurrence = yes for which the classifier predicted credit risk_occurrence = no). These terms are useful when analyzing a classifier's ability and are summarized in Table [3].

Table 3: A confusion matrix for positive and negative tuples

Predicted class		
Actual class	(Credit risk=YES)	(Credit risk=NO)
(Credit risk= YES)	True positives	False negatives
(Credit risk=NO)	False positives	True negatives

The Table [3] shows the confusion matrix for the risky/un-risky predictions. The correctly classified instances are represented as correlation coefficient for the classifier Table [4] and for 48 hour prior prediction in Table [5] which contains the true positive rate (TP Rate), false positive rate (FP Rate), precision, recall, F-measure and ROC area details. The precision is a measure of exactness (i.e., what percentage of tuples are actually labeled as positive), where recall is a measure of completeness (what percentage of positive tuples are labeled as such). The F measure is the harmonic mean of precision and recall. It gives equal weight to

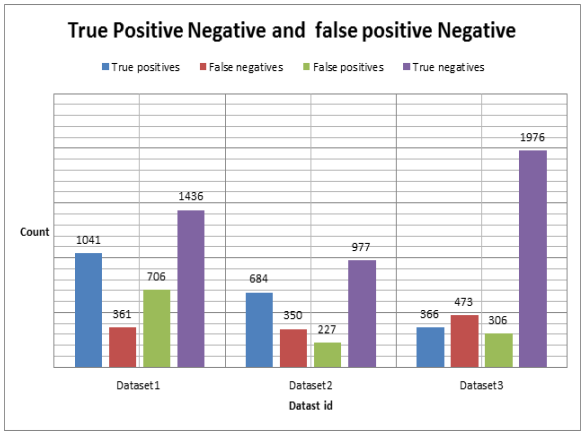
precision and recall. A receiver operating the characteristic curves value describes the trade-off between the true positive rate (TPR) and the false positive rate (FPR). TPR describes the sensitivity and FPR describes the specificity.

$$\text{Precision} = TP / (TP + FP)$$
$$\text{Recall} = TP / (TP + FN)$$

Table 4:Accuracy calculation for the deleted class after using EK-EDT

	Accuracy calculation for the detected class after using EK-EDT					
Dataset id	True positives	False negatives	False positives	True negatives	No. of instances	Accuracy
Dataset1	89	5	5	11	100	99.9 %
Dataset2	109	2	1	8	120	97.5 %
Dataset3	139	3	2	4	150	96%

Table 5:True positive Negative and false positive Negative



CONCLUSION

Banking credit risk analysis has become a difficult task to learn, when the user doesn't know anything about the customer in prior. This toughness is due to the high dimensional and dynamic data. And the loan repayment option differs from customer to customer. This thesis presented an overview of the banking customer's data required for decision making and provides an interactive GUI based tool to analyze the customer type, loan repayment ability and risk score. The decisions related to the customer loan repayment activities are highlighted.

The paper also includes the enhanced EDT algorithm for fast decision support process. This finds the appropriate solutions and decisions, based on the given attributes and values. The experiments are conducted with various conditions and factors to evaluate the output of the proposed system. And the experiments are carried out using Dotnet technology. This helps the organizations in making the right decision to approve or to reject the loan request of the customer by analyzing the credit risk. A Decision Tree Algorithm named as EDT is used for the prediction. And also we obtained the statistically significant linear and nonlinear models to accomplish the above results.

This study also proposed a new classification and prediction scheme for the bank data. And studied the main two problems of the literature, which are prediction accuracy and classification delay. The study overcomes the above two problem by applying the effective enhanced decision tree techniques with data suggestions. The EDT represents the effective splitting criterion which has been verified by the data suggestion. Here, pre pruning and post pruning to eliminate the irrelevant results is also performed. This system effectively identifies the disease and its sub types. The sub type is referred as the percentage of class that are normal and disease.

The experimental results are evaluated using the C#.net. The experimental result shows that the integrated extended decision tree with data suggestion shows better quality assessment compared with the traditional C4.5 techniques. From the experimental results, the execution time calculated for the classification object is almost reduced than the existing system.

REFERENCES

[1] Albayrak A. S., Yılmaz Ş.K., "Data mining: Decision tree algorithms and an application on data of IMKB", SüleymanDemirel University the Journal of Faculty of Economics and Administrative Sciences, vol. 14, pp. 31-52, 2009

[2] Aşan Z., "Examining the socioeconomic characteristics of customers using credit cards, with clustering analysis", Dumlupınar University

The Journal of Social Sciences, vol.17, pp. 256-267, 2007.

[3] Atbaş A. C., A study on determining the cluster number in clustering analysis, Master Thesis, Ankara University, Graduate School of Natural Sciences, 2008.

[4] Bilen H., Data mining application for personnel selection and performance evaluation in banking sector, Master Thesis, Gazi University, Graduate School of Natural and Applied Sciences , 2009.

[5] Chien C.-F., Chen L.-F., "Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry", Expert Systems with Applications,vol. 34, pp. 280-290, 2008.

[6] Ching W. K., Pong M. K.., Advances in data mining and modeling, 1st ed., World Scientific, Hong Kong, China, 2002.

[7] S. Kotsiantis, D. Kanellopoulos, P. Pintelas, "Data Pre-processing for Supervised Leaning", International Journal of Computer Science, 2006, Vol 1 N. 2, pp 111–117

[8] Bharati M. Ramageri, "DATA MINING TECHNIQUES AND APPLICATIONS", Indian Journal of Computer Science and Engineering Vol. 1 No. 4

[9] VivekBhambri "Application of Data Mining in Banking Sector", International Journal of Computer Science and Technology Vol. 2, Issue 2, June 2011

[10] P.Sundari, and Dr.K.Thangadurai "An Empirical Study on Data Mining Applications", Global Journal of Computer Science and Technology, Vol. 10 Issue 5 Ver. 1.0 July 2010.

[11] Kazi Imran Moin, Dr. QaziBaseer Ahmed "Use of Data Mining in Banking",International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622,vol.2,Issue2,Mar-Apr2012, pp.738-742 738

[12] RajanishDass, "Data Mining in Banking and Finance: A Note for Bankers", Indian Institute of Management Ahmadabad.

[13] Hamid EslamiNosratabadi and Ahmad Nadali,"A New Approach for Labeling the Class of Bank Credit Customers via Classification Method in Data Mining", International Journal of Information and Education Technology, Vol. 1, No. 2, June 2011

[14] Hamid EslamiNosratabadi and Ahmad Nadali ,"A New Approach for Labeling the Class of Bank Credit Customers via Classification Method in Data Mining", International Journal of Information and Education Technology, Vol. 1, No. 2, June 2011

[15] Costa, G., F. Folino and R. Ortale, 2007. Data mining for effective risk analysis in a bank intelligence scenario.Preccedings of the 23rd International Conference on Data Engineering Workshop, Apr