Classification approach based Customer Prediction Analysis for Loan Preferences of Customers

Priyanka L.T
Assistant Professor, CSE
Sri shakthi institute of engg & tech, Coimbatore

Neethu Baby
P G scholar, CSE
Sri shakthi institute of engg & tech, coimbatore

ABSTRACT

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 behaviours and prediction. The classifier will predict the customers belongs to which class that should have highest posterior probability. The valuable customer information accumulated by commercial banks, which is used to identify customers and provide decision support. The data pre-processing techniques like data cleaning and data reduction can be applied for data preparation and the dates were converted into a numerical form. A data model is generated based upon the history of the customers in the bank. Then the sample data is classified by using the Naïve Bayesian classification algorithm and placed them into the appropriate class based upon the posterior probability and based upon the posterior probability the percentage of loan sanction risk for the customers can be predicted.

General Terms

Naïve Bayesian Classification algorithm

Keywords

CRM, Data Cleaning, Data Pre-processing, Data Reduction, Naive Bayesian Classification

1. INTRODUCTION

Due to high competition in the business field, it is essential to consider the customer relationship management [1] of the enterprise. Here analyse the massive volume of customer data and classify them based on the customer behaviours and prediction. Customer relationship management is mainly used in sales forecasting and banking areas. Data mining provides the technology to analyse mass volume of data and/or detect hidden patterns in data to convert raw data into valuable information.

Data mining has attracted a great deal of attention in the information industry and in society as a whole in recent years, due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. The information and knowledge gained can be used for applications ranging from market analysis, fraud detection, and customer retention, to production control and science exploration.

Data mining is a step in the knowledge discovery process consisting of particular data mining algorithms. It is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining is primarily used today by companies with a strong consumer focus retail, financial, communication, and marketing organizations. It enables these companies to determine relationships among "internal" factors such as price, product positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics. And, it enables them to determine the impact on sales, customer satisfaction, and corporate profits.

Data mining consists of five major elements:

- Extract, transform, and load transaction data onto the data warehouse system.
- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table.

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 to the posterior probability. Here the data mining concept is used to perform the classification and prediction of loan [3].

Data mining commonly involves four classes of tasks:

- **Clustering** - is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
- **Classification** - is the task of generalizing known structure to apply to new data. For example, an email program might attempt to classify an email as legitimate or spam. Common algorithms include decision tree learning, nearest neighbour, naive Bayesian classification, neural networks and support vector machines.
- **Regression** - Attempts to find a function which models the data with the least error.
- **Association rule learning** - Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.

2. RELATED WORK

Here initially need to create the account for each customers in the bank and they should enter their personal details, income details, insurance details, loan details and the account information of the corresponding customer in other banks. The validation and authentication of the customers were done by the bank manually. Need to store these details in the database for further accessing and for making decisions.

After creating the account for each customer need to prepare the details of customer for data mining. Before performing data mining need to perform the processes like data preparation and data cleaning. In data preparation module the details collected from the customers are converted into a format that is suitable for data mining. In this module the data reprocessing techniques like data cleaning and data reduction were applied for conversion. In data preparation need to select only the wanted fields from each table in order to perform the data mining. After this combine the needed fields into a common table and convert all continuous data into numerical data .In data cleaning need to remove the noise data from the common table.

Data cleaning [4] procedure is used to clean the data by filling the missing values, smoothing noisy data, identifying or removing outliers and resolving inconsistencies. If the user is believe that the data are dirty, and then they will not trust the results of the data mining process that has been applied to this data.Furthermore, dirty data can cause confusion for the mining procedure, resulting in unreliable output. Although most mining routines have some procedures for dealing with incomplete or noisy data, they are not always robust. Instead, they may concentrate on avoiding over fitting the data to the function being modeled. Therefore, before performing the data mining we need to run data through some data cleaning routines. [2]

In addition to data cleaning, step must be taken to help and avoid redundancies during data integration. Typically, data cleaning and data integration are performed as a data preprocessing step when preparing for a data mining. Additional data cleaning can be performed to detect and remove redundancies still occur in the results obtained after data integration.

The customer data may contain certain attribute that will take larger values. Therefore if the attributes are left UN normalized, we need to normalize that. Furthermore, it would be useful for analysis to obtain aggregate information. The data transformation operations, such as normalization and aggregation, are additional data preprocessing procedures that would contribute toward the success of the mining process.

Data reduction produces a reduced representation of the data set that is much smaller in volume and that should produce the same result. There are many methods used for data reduction they are data aggregation, attribute subset selection, dimensionality reduction and numerosity reduction. In data aggregation made a data cube corresponding to data. Attribute subset selection is used to remove the irrelevant attributes from table through correlation analysis. Dimensionality reduction makes use of encoding schemes such as minimum length encoding or wavelets encoding. In numerosity reduction, replacing the data with alternate or smaller representations such as clusters or parametric models.

In bank it is necessary to analyze the customer data in order to learn which loan applicants are safe and which are risky for the bank. The process of analyze the data is known as data classification, here a model or classifier is constructed to predict categorical labels such as safe or risky for the loan application data.

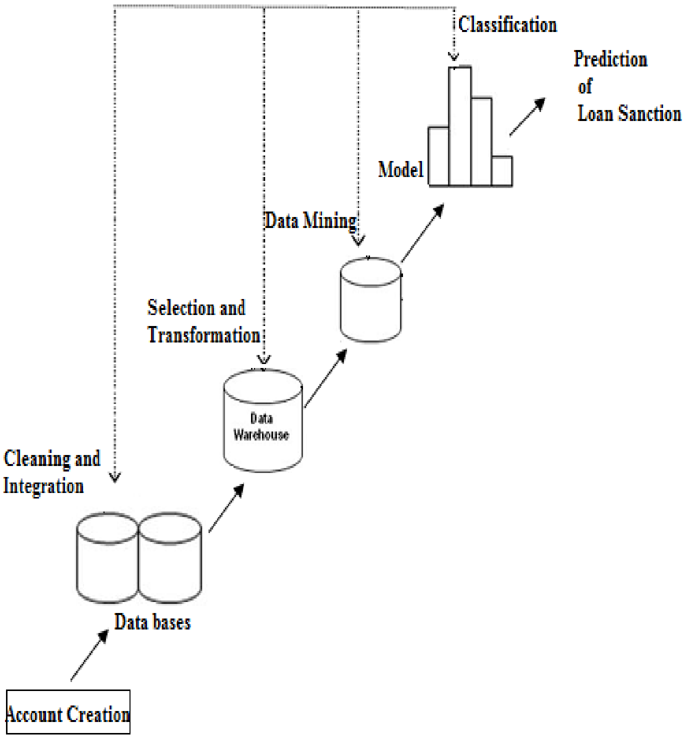


Fig-1.Block Diagram

3. CLASSIFICATION AND PREDICTION

Data classification is a two step process. First step is the learning process. In this the training data are analyzed by a classification algorithm here, the class label attribute is loan decision. The second process is classification; here test data are used to estimate the accuracy of the classification algorithm. If the accuracy of the classification is better than that algorithm can be applied to the classification of new data tuples.

Data prediction is a two-step process, similar to that of data classification. However, for prediction, we lose the terminology of “class label attribute” because the attribute for which values are being predicted is continuous-valued (ordered) rather than categorical (discrete-valued and unordered). The attribute can be referred to simply as the predicted attribute. The accuracy of a predictor is estimated by computing an error based on the difference between the predicted value and the actual known value of y for each of the test tuples, $\boldsymbol{X}$.

3.1Naive Bayesian Classifier Algorithm

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class. Bayesian classification is based on Bayes’ theorem.

Naïve Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence. It is made to simplify the computations involved and, in this sense, is considered “naïve.” Bayesian belief networks are graphical models, which unlike naïve Bayesian classifiers allow the representation of dependencies among subsets of attributes. Bayesian belief networks can also be used for classification.

This algorithm will works as follows

Consider let *D* be a training set of tuples and their associated class labels. As usual, each tuple is represented by an *n*-dimensional attribute vector, ***X*** = (*x*1, *x*2, : : : , *xn*), here *n* measurements made on the tuple from *n* attributes, respectively, *A*1, *A*2, ….., *An*.

Consider there are *m* classes, *C*1, *C*2, ……, *Cm*. Given a tuple, ***X*** need to be classified and predict, the classifier will predict that ***X*** belongs to the class having the highest posterior probability, conditioned on ***X***. That is, the naïve Bayesian classifier predicts that tuple ***X*** belongs to the class *Ci* if and only if

$$P(C_i|X) > P(C_j|X) \quad \text{for } 1 \leq j \leq m, j \neq i. \qquad \text{-(1)}$$

Then we need to maximize the value of P (Ci|X). The class Ci for which P (Cij**X**) is maximized is called the maximum posteriori hypothesis. By Bayes' theorem

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}. \qquad \text{-(2)}$$

As P(**X**) is constant for all classes, only P (**X**jCi)P(Ci) need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, P(C1) = P(C2) = _ _ _ = P(Cm), and we would therefore maximize *P*(***X***j*Ci*). Otherwise, we maximize *P* (***X***j*Ci*) *P* (*Ci*). Given data sets with many attributes, it would be extremely computationally expensive to compute *P* (***X***j*Ci*). In order to reduce computation in evaluating *P* (***X***j*Ci*), the naive assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the tuple.

$$\begin{aligned} P(X|C_i) &= \prod_{k=1}^{n} P(x_k|C_i) \\ &= P(x_1|C_i) \times P(x_2|C_i) \times \cdots \times P(x_n|C_i). \end{aligned}$$

We can easily estimate the probabilities *P* (*x*1|*Ci*), *P*(*x*2|*Ci*), … , *P*(*xn*|*Ci*) from the training tuples.

Here in this example I have consider only few fields similarly we need to consider many fields to predict the loan sanction.

Let C1 correspond to the class of loan sanction=' yes ' and C2 correspond to the class of loan sanction='no'. We need to classify and the test data

X=(acctype=SB,age=40,tax=yes,customer,type=staff,qualification=U.G,income=6lakhs)

We need to maximize *P*(*X Ci*)*P*(*Ci*) , for *i* = 1,2 ,3...*P*(*Ci*) , the prior probability of each class, can be computed based on the training samples:

P (*Loan sanction* =" *yes*") = 8 /15 = 0.533

P (*Loan sanction*="*no*") = 7 /15 = 0.466

To compute *P*(*X Ci*) , for *i* = 1,2 , we compute the following conditional probabilities: The posterior probability of account type can be calculated as

P(acctype ="SB" Loan sanction=" yes")=4/8=0.5

P(acctype ="SB" Loan sanction=" no")=3/7=0.428

The posterior probability of age can be calculated as following

P (age ="40" Loan sanction=" yes") =7/8=0.875

P (age ="40" Loan sanction=" no") =5/7=0.714

The posterior probability of tax can be calculated as

P (Tax ="yes" Loan sanction=" yes") = 7/8=0.875

P (Tax="yes" Loan sanction=" no") =5/7=0.714

The posterior probability of customer type can be calculated as

Table-I: ***Training Data Of Customer Stored In Database***

Id	**Acctype**	**Age**	**Tax**	**Custo-mer type**	**Qual-ification**	**Income**	**Loan Sanction**
1	SB	18-60	No	Staff	P.G	>10lakhs	Yes
2	SB	18-60	Yes	Public	PLUS TWO	1-5 lakhs	No
3	SB	>60	Yes	Public	P.G	1-5 lakhs	Yes
4	FD	18-60	Yes	Staff	PLUS TWO	5-10 lakhs	Yes
5	FD	18-60	Yes	Public	U.G	1-5 lakhs	Yes
6	FD	18-60	Yes	Staff	U.G	<1 lakh	No
7	FD	18-60	Yes	Staff	P.G	<1 lakh	No
8	FD	18-60	Yes	Staff	U.G	5-10 lakhs	Yes
9	SB	18-60	Yes	Staff	P.G	>10 lakhs	Yes
10	CA	18-60	Yes	Staff	PLUS TWO	<1 lakh	No
11	FD	18-60	Yes	Public	P.G	5-10 lakhs	Yes
12	SB	>60	No	Public	$<10^{TH}$	<1 lakh	No
13	CA	18-60	Yes	Public	$<10^{TH}$	1-5 lakhs	No
14	SB	18-60	Yes	Staff	P.G	>10 lakhs	Yes
15	SB	>60	No	Public	<10TH	<1 lakh	No

P (Customer Type ="staff" Loan sanction=" yes") =5/8=0.625

P (Customer Type=”staff” Loan sanction=" no") =3/7=0.428

The posterior probability of qualification can be calculated as following

P (Qualification ="U.G" Loan sanction=" yes") =2/8=0.25

P (Qualification ="U.G" Loan sanction=" no") =1/7=0.142

The posterior probability of income can be calculated as following

P (income ="6 lakhs" Loan sanction=" yes")=7/8=0.875

P (income ="6 lakhs" Loan sanction=" no")=5/7=0.714

Using the above probabilities, we obtain

P (*X Loan sanction* =" *yes*") =
0.5*0.875*0.875*0.625*0.25*0.875=0.0523

P(*X Loan sanction* ="*no*")=
0.428*0.714*0.714*0.428*0.142*0.714=0.0094

The total probability can be calculated as

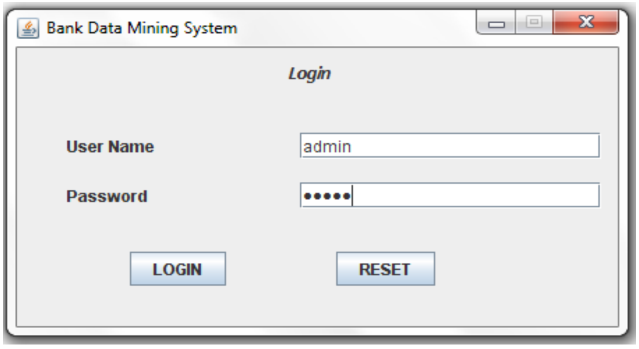
P(*X Loan sanction* =" *yes*")*P*(*Loan sanction* ="*yes*")=
0.0523*0.533 = 0.0278

P(*X Loan sanction* ="*no*")*P*(*Loan sanction* ="*no*")= 0.0094*
0.466 = 0.0043

Here the probability of loan_santion=”yes” is greater than Loan sanction=”no”.hence here we predict, the loan can be sanctioned to that particular customer.

4. CONCLUSION AND RESULTS

Here analyse the massive volume of customer data and classify them based on the customer behaviours and prediction. Customer relationship management is mainly used in sales forecasting and banking areas. Data mining provides the technology to analyse mass volume of data and/or detect hidden patterns in data to convert raw data into valuable information. The classifier will predict the customers belongs to which class that should have highest posterior probability. The valuable customer information accumulated by commercial banks, which is used to identify customers and provide decision support. The details collected from the customers are converted into a format that is suitable for data mining and is called data preparation. The data pre-processing techniques like data cleaning and data reduction can be applied for data preparation and the data were converted into a numerical form. A data model is generated based upon the history of the customers in the bank. Then the sample data is classified by using the Naïve Bayesian classification algorithm.



User Name: admin
Password: •••••

Fig-2.Login window

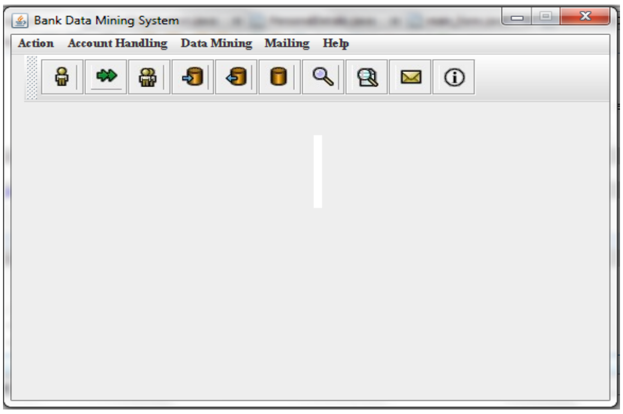
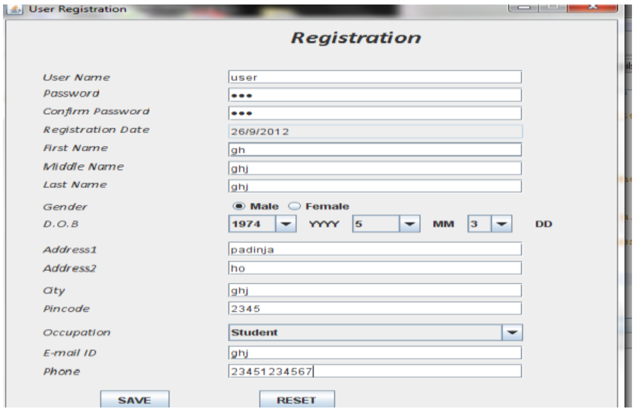
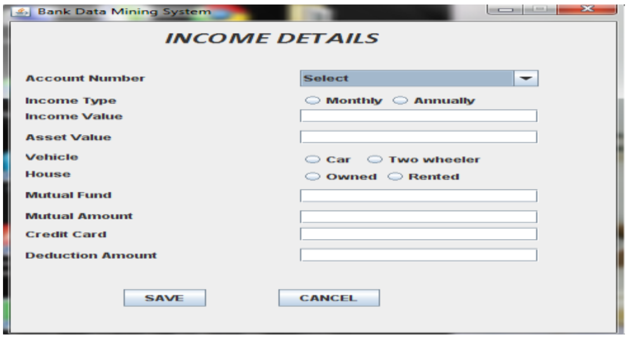


Fig-3.Main window



User Name: user
Password: •••
Confirm Password: •••
Registration Date: 26/9/2012
First Name: gh
Middle Name: ghj
Last Name: ghj
Gender: ● Male ○ Female
D.O.B: 1974 YYYY 5 MM 3 DD
Address1: padinja
Address2: ho
City: ghj
Pincode: 2345
Occupation: Student
E-mail ID: ghj
Phone: 23451234567

Fig-4.Registration window


Bank Data Mining System

INCOME DETAILS

Account Number: Select

Income Type: ○ Monthly ○ Annually

Income Value

Asset Value

Vehicle: ○ Car ○ Two wheeler

House: ○ Owned ○ Rented

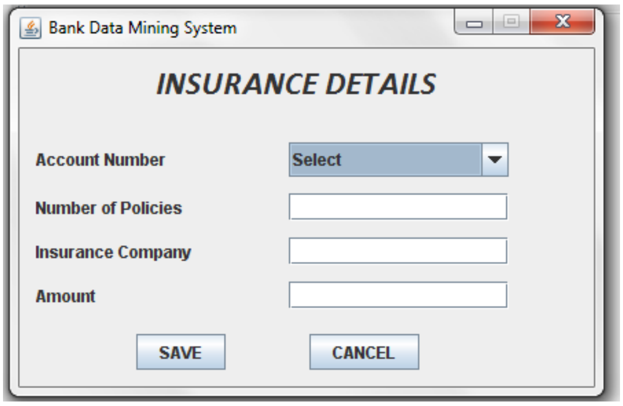
Mutual Fund

Mutual Amount

Credit Card

Deduction Amount

Fig-5.Income details window


Bank Data Mining System

INSURANCE DETAILS

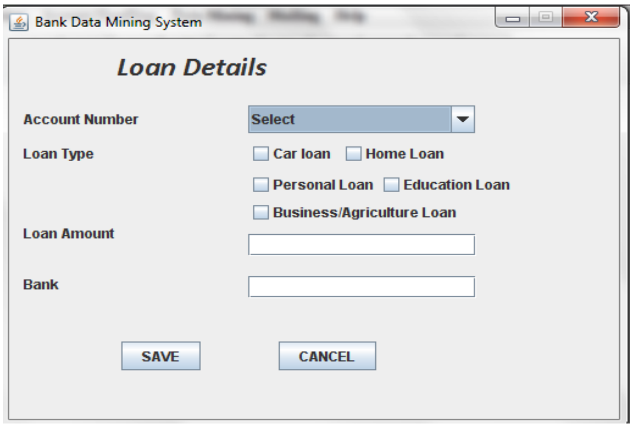
Account Number: Select

Number of Policies

Insurance Company

Amount

Fig-6.Insurance details window


Bank Data Mining System

Loan Details

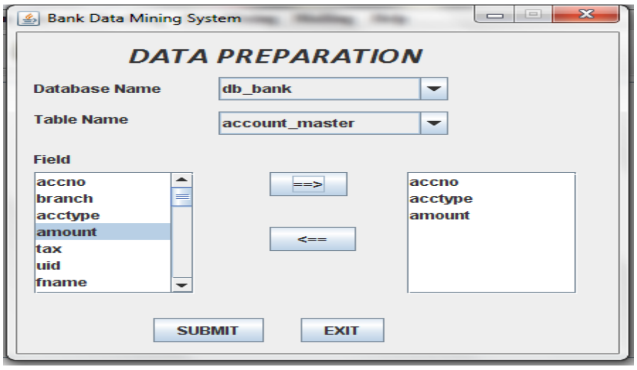
Account Number: Select

Loan Type: □ Car loan □ Home Loan □ Personal Loan □ Education Loan □ Business/Agriculture Loan

Loan Amount

Bank

Fig-7.Loan details window


Bank Data Mining System

DATA PREPARATION

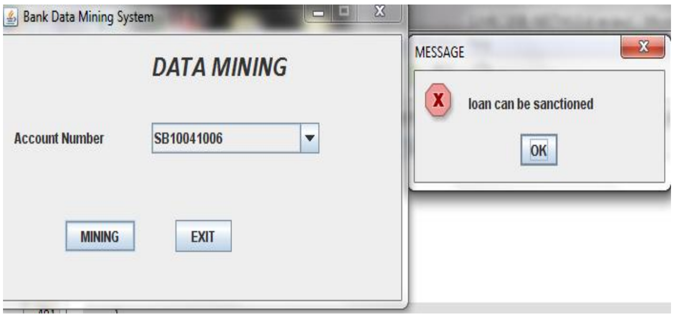
Database Name: db_bank

Table Name: account_master

Field: accno, branch, acctype, amount, tax, uid, fname

accno, acctype, amount

Fig- 8. Datapreparation window


Bank Data Mining System

DATA MINING

Account Number: SB10041006

MESSAGE

loan can be sanctioned

Fig-9. Datamining window

5.REFERENCES

[1] Jenkins D (1999). “Customer relationship management and the data warehouse” [J]. Call center Solutions, 1892):88-92

[2] Chung HM and P Gray(1999). “Data mining” [J]. Journal of MIS,16(1):11-13

[3] HOKEY MIN, Developing the Profiles of Supermarket Customers through Data Mining[J], The Service Industries Journal, Vol.26, No.7, October 2006, pp.747–763

[4] Balaji Padmanabhan, Alexander Tuzhilin , On the Use of Optimization for Data Mining: Theoretical Interactions and eCRM Opportunities[J], Management Science . Vol. 49, No. 10, October 2003, pp. 1327 1343

[5] Davis B(1999). “Data mining transformed” [J]. Informationweek, 751:86-88

[6] Kuykendall L(1999).”The data-mining toolbox”[J]. Credit Card Management,12(6):30-40

[7] Han, J. and Kamber M., (2001) Data Mining: Concepts and Techniques[M], San Francisco, CA: Morgan Kaufmann Publishers