Rule based Model for Credit Evaluation using Rough Set Approach

Anjali Kulkarni
UDCS and Dept. of Computer Science,
CKT College of Arts, Science and Commerce,
University of Mumbai, India,
Email: anjali_kulkarni74@rediffmail.com

Dr. Seema Purohit
Kirti College of Arts, Science and Commerce
University of Mumbai, India,
Email:supurohit@gmail.com

Abstract: Credit plays an important role in the economy. Credit Evaluation of any potential credit application has remained a challenge for banks all over the world till today. Credit evaluation can be defined as a technique that helps lenders to decide whether to grant credit to consumers or not. Its increasing importance can be seen from the growing popularity and application of credit scoring. It is mandatory not only to construct effective credit scoring models to help improve the bottom-line of credit providers, but also to design rule based system for effective credit evaluation system. This paper approaches the use of rough set technique to generate rule based system for credit scoring model.

Keywords: *Credit evaluation, Rough Set Techniques, Exhaustive Algorithm, Rule Based Model*

*****

I. INTRODUCTION

A. *Credit Evaluation*

The process of Credit Evaluation is called before the credit score is sanctioned which may take time, but concludes in either an approval or a rejection.

Credit approval process is a potential application as it is related to the economy of a country. There are five criteria's involved in credit evaluation which are: character, credit report, capacity, cash flow, and collateral.

A person with a sound financial objective is likely to be granted a credit approval quickly than an individual who is in bad shape, not just on the financial facet, but also on other aspects.

Credit history is another important factor considered by lenders in their decision to grant and approve credit applications. The credit report is a record of an individual's past borrowing and reimbursing transactions. It also includes information about late payments and bankruptcy [1].

A credit report can be tarnished. If the cash flow is good, then there is a possibility to grant credit approval. Lenders may also have to check the liquidity of an individual. This can be done by checking the bank statements of an individual borrower. In the case of businesses, lenders may have to obtain a copy of the audited financial statements.

The financial statements of businesses and bank statements can be utilized to show the capacity of a borrower to settle and repay a line of credit. The capacity of the borrower to pay a credit is determined during credit evaluation and approval.

A lender seeks for security whenever the borrower defaults the credit payment. If no collateral is present as security for a credit, it is likely that the lender will give the borrower a high-interest rate credit.

B. *Decision process for Credit Evaluation*

A credit manager must evaluate the risk associated with extending credit and declining an applicant based on numerous factors.

The need for sufficient and reliable information is the foundation of a successful credit decision. A credit manager may call on references, run background checks, pull a credit report, verify bank accounts or ask questions of the applicant to validate the information on the credit application. Credit managers are challenged with the task of obtaining readily available information to support their decision while sending a timely response to the applicant. A major obstacle in achieving this task is the turnaround time associated with checking references. The process varies from business to business and may include a background check, a verification of a bank deposit or credit references with existing suppliers. Some businesses require written requests, while others may offer to do a phone interview at their convenience [1].

The lack of general credit review system in many banks and the lack of precise methods for measuring credit risk are two important reasons why an expert support system is necessary. Such a system can be implemented by using Rough set approach as well as integrated approach. The integrated model can be generated using the combination of Logistic Regression, Support Vector Machine, Radial Basis Function and Decision Tree where out of four methods one is top classifier and remaining are base classifiers. Rough set generates rule based model which extracts valuable information from database and though it takes little time it is possible to generate integrated rule based model of Logistic Regression, Support Vector Machine, Radial Basis Function and Decision Tree using rough set.

C. *Rough Set Approach*

Rough Set Theory (RST). Rough set theory, first proposed by Pawlak [8] in 1980s, employed mathematical modeling to deal with class data classification problems, and then turned out to be a very useful tool for decision support systems, especially when hybrid data, vague concepts and uncertain data were involved in the decision process. The rough set theory offers a viable approach for decision rule extraction from data.

The big problem in data mining is the deficiency and indeterminateness. This problem is solved by using new theories and procedures, rough sets. It offers the mathematic tools for discovering hidden patterns in data through the use of identification of partial and total dependencies in data. It also enables work with null or missing values. The rough sets theory uses different approach to uncertainty. The advantage of this method is a mathematic base of rough sets and the possibility of mathematic description of this problem. Rough sets seem to be advantageous for mining of incomplete information as well as for other algorithms.

The main goal of the rough set analysis is induction of (learning) approximations of concepts. Rough sets constitute a sound basis for KDD. It offers mathematical tools to discover patterns hidden in data. It can be used for feature selection, feature extraction, data reduction, decision rule generation, and pattern extraction (templates, association rules) etc. It identifies partial or total dependencies in data, eliminates redundant data and gives approach to null values, missing data, dynamic data.

Advantages of Rough Set:

1] Provides efficient algorithms for finding hidden patterns in data.

2] Identifies relationships that would not be found using statistical methods.

3] Allows both qualitative and quantitative data

4] Finds minimal sets of data (data reduction).

5] Evaluates significance of data.

6] Generates sets of decision rules from data.

7] It is easy to understand.

8] Offers straightforward interpretation of obtained results.

To use the rough set process, one begins with a relational database, a table of objects with attributes, and attributes values for each object. One attribute is chosen as the decision attribute, then the rest of the attributes are the condition attributes [8].

The RST approach is based on refusing certain set boundaries, implying that every set will be roughly defined using a lower and an upper approximation [9].

For example, let B ∈ A and X ∈ U be an information system. The set X is approximated using information contained in B by constructing lower and upper approximation sets, respectively: $\underline{B}X = \{xj[x]B \in X\}$ (lower) and $BX = \{xj[x]B \cap X \neq \emptyset\}$ (upper). The elements in BX can be classified as members of X by the knowledge in B. The set $BNB(x) = BX- \underline{B}X$ is called the B-boundary region of X and it consists of those objects that cannot be classified with certainty as members of X with the knowledge in B. The set X is called `rough' with respect to the knowledge in B if the boundary region is non-empty. Rough sets theoretic classifiers usually apply the concept of rough sets to reduce the number of attributes in a decision table [2] and to extract valid data from inconsistent decision tables. Rough sets also accept discretized (symbolic) input.

The structure of data is represented in the form of information system called decision table. An information system is a pair S = (U, A) where U is a non-empty, finite set of objects and A is a non-empty, finite set, of attributes. Each $a \in A$ corresponds to the function a: $U \rightarrow Va$ called evaluation function, where Va is called the value set of a. In information table, usually rows associated to its objects, its columns to attributes and its cells to values of attributes on objects. In supervised learning problems, objects from training set are pre-classified into some categories or classes.

D. Significance and Objective of study

The main requirement of assessing the loan applications received from the borrowers is having a mechanism of credit evaluation which will reduce the potential risk of credit applications. The scalability of the credit evaluation system should be good enough to handle a large volume of credit applications quickly with minimal labor, thus reducing operating costs. These systems may be an effective substitute for the use of judgment among inexperienced credit officers, thus helping them to control bad debt losses. Hence if such a full proof system is designed it will take care of the risk involved in the credit evaluation process. Use of Rough sets generates a Rule based systems having a strong mathematical base and hence they are more accurate. They generate the rules by which the entry of a wrongful application in the system is prevented. Hence the objective of the study is to design the rule based model which can be constructed by using rough set technique such that there will be minimum defaulters and credit risks.

II. PROBLEM DEFINITION

The banking data required for Loan Approval process is usually huge and the data mining classification techniques gives the status of individual application received. Further if we want to generate the rules governing the classification so that the percentage of misclassification will be reduced. Hence there is a need to design a general rule based precise credit evaluation and risk measuring system.

III. RULE BASED CREDIT EVALUATION MODEL

A. Requirements and Architecture

Credit Evaluation System requires the following four components to work with data: Computer System, RSES 2.2 Software, Customer, and Data. Rule based Credit evaluation system can be built by rough set reduct technique.

B. Identification of Independent and Dependent Variables

The data set used in this research is divided into training and testing data sets. All training cases are set by default taking into account the banks' guidelines for personal credit approval in the banks. 45000 customer's data for the current study was collected from different banks such as SBI, IDBI, AXIS and Syndicate banks. It consists of different independent variables and one dependent variable.

Variables are the conditions or characteristics that the investigator manipulates, controls or observes. Variables are classified as dependent and independent variables. An independent variable is the condition or characteristic that affects one or more dependent variables: its size, number, length or whatever exists independently and is not affected by the other variable. A dependent variable changes as a result of changes to the independent variable.

Independent Variables (for 45000 customer's data)

1) Age
2) Job
3) Marital Status

4) Education
5) Defaulter
6) Balance
7) Housing Loan
8) Other loan

Dependent Variable:
1) Credit (Approved or Not)

IV. TECHNOLOGY

A. *Software Customization*

All the experiments are carried out through RSES (Rough Set Exploration System. Rough Set Exploration System (RSES) is a software tool designed and implemented at Warsaw University. It is a library of methods and a graphical user interface supporting variety of rough set-based computations.

B. *Methodology using RSES*

Rule extraction techniques from data

Rules can be generated using following algorithms:

1. Exhaustive Algorithm: This algorithm realizes the computation of object oriented reducts (or local reducts). It has been shown that any minimal consistent decision rules for a given decision table S can be obtained from objects by reduction of redundant descriptors. The method is based on Boolean reasoning approach [3].

The algorithm is given below:

```
Exhaustive (int sol, int depth)
{
if
(issolution (sol))
printsolution (sol)
else
{
solgenerated=generatesolution()
exhaustive (solgenerated, depth+1)
}
}
```

2. Genetic Algorithm: Using genetic algorithm with permutation encoding and special crossover operator [3], one can compute a predefined number of minimal consistent rules.

3. Covering Algorithm: This algorithm searches for minimal (or very close to minimal) set of rules which cover the whole set of objects [3].

The algorithm is given below:

```
Inputs: labeled training dataset D
Outputs: Rule set R that covers all instances in D
Procedure:
Initialize R as the empty set
for each class C
{
while D is nonempty
{
Construct one rule r that correctly classifies some
instances in D that belong to class C and does not
incorrectly classify any non-C instances
Add rule r to rule set R
Remove from D all instances correctly classified by r
}
}
Return R
```

4. LEM2 Algorithm: This is another kind of covering algorithm strength of a rule, based on the measure of support, consistency, and coverage [3].

```
Input: B set of objects
Output: R set of rules
begin
G=B;
R=φ;
While G≠φ
do
begin
  C≠φ
  C (G) ={c: [c] ∩G≠φ};
  While (C≠φ) or (! ([C] ⊆B))
  do
  begin
  select a pair c∈C (G) such that |[c]∩G| is maximum;
  if ties, select a pair c ∈C (G) with the smallest
  cardinality |[c]|;
  if further ties occur, select the first pair from the list;
  C=C∪{c};
  G=[c] ∩G;
  C (G) ={c: [c] ∩G≠φ};
  C (G) =C (G)–C;
  end;
  for each elementary condition c∈C
  do
  create rule r basing the conjunction C and add it to R;
  G==B–U|R|
  r∈R
  end;
  for each r ∈R
  do
  if U|S|=B then R=R–r
  S∈R-r
  end
```

V. RESEARCH FINDINGS

Table I. Rules generated using Exhaustive Algorithm

S. No.	Rules
1	("age "=33.0)&(" job "=entrepreneur)&("default "=no)=>("Credit Risk"={low})
2	("age "=47.0)&(" job "=blue-collar)&("default "=no)=>("Credit Risk"={low})
3	("age "=35.0)&(" job "=management)&("default "=no)=>("Credit Risk"={low})
4	("age "=28.0)&(" job "=management)&("default "=no)=>("Credit Risk"={low})
5	("age "=42.0)&(" job "=entrepreneur)&("default "=yes)=>("Credit Risk"={high})
6	("age "=58.0)&(" job "=retired)&("default "=no)=>("Credit Risk"={high})

Table II. Rules generated using Genetic Algorithm

S. No.	Rules
1	("age "=67.0)&(" job "=blue-collar)&("default "=no)=>("Credit Risk"={high})
2	("age "=23.0)&(" job "=unemployed)&("default "=no)=>("Credit Risk"={high})
3	("age "=93.0)&(" job "=retired)&("default "=no)=>("Credit Risk"={high})
4	("age "=30.0)&(" job "=lecturer)&("default "=no)=>("Credit Risk"={low})
5	("age "=26.0)&(" job "=self-employed)&("default "=no)=>("Credit Risk"={low})

Table III. Rules generated using Covering Algorithm

S. No	Rules
1	("age "=74.0)=>("Credit Risk"={high})
2	("age "=19.0)=>("Credit Risk"={high})
3	("age "=68.0)=>("Credit Risk"={high})
4	(" job "=retired)=>("Credit Risk"={high})
5	(" job "=unemployed)=>("Credit Risk"={high})
6	("default "=yes)=>("Credit Risk"={high})

Table IV. Rules generated using LEM2 Algorithm

S. No	Rules
1	("loan "=no)&("default "=no)&(" housing "=no)&(" marital "=married)&(" job "=retired)=>("Credit Risk"={high})
2	("default "=no)&("loan "=no)&(" balance "='\'(13.5-1839.5]\'")&(" housing "=yes)&(" marital "=single)&("education "=tertiary)&(" job "=management)=>("Credit Risk"={low})
3	("default "=no)&(" housing "=yes)&(" balance "='\'(13.5-1839.5]\'")&("loan "=yes)&(" marital "=married)&("education "=secondary)&(" job "=blue-collar)=>("Credit Risk"={low})
4	("default "=no)&("loan "=no)&(" marital "=married)&(" housing "=yes)&("education "=secondary)&(" balance "='\'(1839.5-inf)\'")&(" job "=blue-collar)=>("Credit Risk"={low})

VI. CONCLUSION AND FUTURE WORKS

Research findings indicate that in case of rule based model it achieves more accuracy because in this model each and every record will be visited and accordingly best rules will be generated. It also reduces the risk and misclassifications within the data. So the objective of finding the accurate rule based model is achieved. In future 1) The hybridized model of rough set and SVM can be used to solve problem of risk organization and also 2) The integrated model of SVM, LR, RBF and DT solves the problem of Credit Evaluation.

ACKNOWLEDGMENT

In this study many people and Institution assisted us and without their help this study would not have been completed. A. N. Kulkarni specially thanks Bank officials of many banks in Maharashtra, who helped by providing the data needed for the study.

REFERENCES

[1] S. U. Purohit, A. N. Kulkarni: "Applicability of Integrated model to improve the Decision Process for Credit Evaluation for Indian Banks",IJMO 2012 v2 - 127, volume 2(4):529-534 ISSN: 2010-3697, DOI: 10.7763/IJMO.2012.V2.127
[2] Z. Pawlak: Rough sets: Theoretical aspects of reasoning about data, Kluwer Dordrecht, 1991.
[3] Rough Set Exploration System (RSES) available at http://logic.mimuw.edu.pl/ rses/.
[4] Z. Pawlak: Rough sets and decision analysis, *INFOR: Information system and operational research*, **38**(3), 2000, 132–144.
[5] Z. Pawlak.: Rough sets and decision algorithms, in Rough Sets and Current Trends in Computing (Second International Conference, RSCTC 2000), Springer, Berlin, *RSCTC*, 2001, 30–45.
[6] Suman Saha, C. A. Murthy, Sankar K. Pal: Rough Set Based Ensemble Classifier. Lecture Notes in Computer Science, Vol. 6743, 27-33, 2011
[7] Hala S. Own, Ajith Abraham: A New Weighted Rough Set Framework Based Classification for Egyptian Neonatal Jaundice. Applied Soft Computing. Elsevier Vol. 12(3), 999-1005, 2012
[8] Z. Pawlak, Rough sets, Informational Journal of Computer and Information Sciences, vol.11, no.5, pp. 341-356, 1982.
[9] M. Inuiguchi and M. Tsurumi, Measures based on upper approximations of rough sets for analysis of attribute importance and interaction, International Journal of Innovative Computing, Information and Control, vol.2, no.1, pp.1-12, 2006.
[10] J. Wr´oblewski, Covering with Reducts - A Fast Algorithm for Rule Generation ,Proceeding of RSCTC'98, LNAI 1424, Springer Verlag, Berlin, 1998, pp. 402-407.\