Attribute Reduction for Credit Evaluation using Rough Set Approach

Anjali Kulkarni
Dept. of Computer Science, UDCS, University of Mumbai
C. K. Thakur College, New Panvel
Navi Mumbai, India
Email:anjali_kulkarni74@rediffmail.com

Dr. Seema Purohit
Department of Mathematics
Kirti College, Dadar(W)
Mumbai, India
E-mail: supurohit@gmail.com

Abstract—Generation of an Integrated Model is an important technique in the research area. It is a powerful technique to improve the accuracy of classifiers. This approach has been applied to different types of real time data. The unprocessed data leads to give wrong results by using some of the machine learning techniques. For generation of an integrated model attribute reduction and re-sampling technique is necessary. For attribute reduction Rough set is the best approach as it requires less execution time, high Interpretability, high reduction rate and high accuracy

Keywords*- Credit evaluation, Rough Set, Exhaustive Algorithm and Attribute Reduction*

*****

I. Introduction

A. Credit Evaluation

In today's highly competitive world the business landscape is changing rapidly. Staying ahead in competition and protecting market share means finding new and more efficient ways of doing business. This is reflected in the increasing demand for objective information to help make sound business decisions at the consumer level. Cash flow holds the key for running a successful business successfully. Further, business houses need huge funding for future expansions/ developments/ diversification. Besides the business houses, even individuals, professionals, entrepreneurs, small and medium scale Industries also require funds for setting up businesses/ working capital needs. They rely on external agencies for the fund borrowing.

As a part of their security policies and mechanism of handling risk; external agencies often see the credibility and creditability of the borrower. Creditability of the borrower depends on many factors e.g. cash flow, collateral, payment history, image of the borrower, Insurance coverage, Risks ... etc. To check creditability of the borrower Credit Evaluation procedure needs to be in place. Credit evaluation is one of the important considerations of financial institutions. It leads to either acceptance or rejection of the borrower's request. In Credit Evaluation a standard procedure is followed to assign the credit score to the borrower and honor his request. The affirmative acceptance of the borrower's request depends upon lenders motivation to lend money and borrowers readiness to return money.

B. Need of Integrated Model

Even after the attribute reduction the classification accuracy remains the same and the misclassification percentage is not reduced. This motivates the need of Integrated Model which improves the accuracy of the model. Different classifiers generally make different assumptions on the same sample of data.

Generation of Integrated Model is the process by which multiple models such as classifiers or experts are strategically combined to solve computational intelligence problem. It is used to improve the performance of an existing model and to assign a confidence to the decision made by the model in selecting optimal features, performing data fusion, and incremental learning etc. Integrated Model can be useful when dealing with large data base or inadequate database.

When the amount of data is too large it is difficult for single classifier to train that large database. But in case of integrated Model this large data set is can be partitioned into smaller subsets. Each partition can then be used to train a separate classifier which can then be combined using an appropriate combination rule.

The three primary considerations for integrated model are statistical, computational and representational. The statistical consideration is related to lack of adequate data to properly represent data distribution. The computational consideration facilitates the process of making a choice of model among many models that can solve a given problem, which often becomes difficult for existing model.

The representational consideration addresses the issues when the chosen model cannot properly solve the given problem.

C. Attribute Reduction

Generally Credit data size is huge having many attributes. Attribute reduction is an important topic of research. So it is necessary to reduce that attribute value so that it achieves higher accuracy, low execution time and reduces complexity.

Following are the methods for attribute reduction:

Table I. Attribute Reduction Methods

	Evaluation Parameters			
Name of attribute reduction method	Execution Time	Loss of Interpretation	Reduction Rate or performance	Column (attribute) size and type
Missing Value Ratio	Low computational time as searching for missing values	Average interpretability	Good reduction rate without compromising performance	Both numeric as well as nominal
Low Variance Factor	Normalization is required so high computational time	Average interpretability	Moderate reduction Rate	Only for numeric columns
High Correlation Factor	High computational time	Average interpretability	Moderate reduction Rate	No correlation available between numeric and nominal columns
Random Forest & Ensemble Trees	High computational time	High interpretability	Higher reduction rate	Both numeric as well as nominal columns
Principle Component Analysis	Normalization is required so high computational time	Low interpretability	Worst in reduction rate and model performance	Only for numeric columns
Backward Feature Elimination and Forward Feature Construction	Long execution time Slow on high dimensional data set	Loses its interpretability	Higher reduction rate and higher accuracy	Low number of input columns
Grid Search Method	High computational time	Interpretability is below average	Performance depends on range of predefined parameters	Low dimensional parameters
Direct search method	Low computational time	Interpretability is below average	Good for low dimension parameters	Low dimensional parameters
Rough Set technique	Low computational time	High interpretability as rules can be generated using rough set	Higher reduction rate and higher accuracy	Both numeric as well as nominal columns

From the above table it is observed that Rough set technique is the best technique for attribute reduction for high dimensional data sets as it requires low computational time, highly interpretable and high attribute reduction rate.

D. *Rough Set Approach*

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 [6].

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 [6] and to extract valid data from inconsistent decision tables. Rough sets also accept discretized (symbolic) input.

E. *Significance and Objective of the study*

Generally for credit evaluation, the banking data used is a huge data set having high dimension attributes. To improve accuracy of credit evaluation integrated technique is the best technique. But before applying integrated technique it is necessary to reduce attributes. If these attributes are reduced then accuracy of the credit evaluation is improved. Hence it is necessary to reduce attribute for high dimensional data set.

For reduction of attributes rough set approach is the best approach as it provides higher accuracy, lower execution time and higher reduction rate for attributes. The objective of the study is that first reduce the attributes and then apply integrated technique to improve the accuracy of model.

II. Data and Reduct Generation

A. *Requirements and Architecture*

The following four components to work with data: Computer System, RSES 2.2 Software, Customer, and Data.

B. *Identification of Independent and Dependent Variables*

The data set used in this research is legacy dataset from UCI repository of 4521 customers which is divided into training and testing data sets. This data set consists of dependent and independent variables. Variables are the conditions or characteristics that the investigator manipulates, controls or observes. 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. The dependent and independent variables of this data set is given as follows:

Independent Variables

1) Age
2) Job
3) Marital Status
4) Education
5) Defaulter Status
6) Balance
7) Housing Loan Status
8) Other loan Status
9) Contact
10) Day
11) Month
12) Duration
13) Campaign
14) Pdays
15) Previous
16) Outcomes

Dependent Variable:

1) Credit (Approved or Not)

III. Methodology

A. *Software Customization*

RSES (Rough Set Exploration System and WEKA is used to analyze bank customer data set. Rough Set Exploration System (RSES) is a software tool designed and implemented at Warsaw University. RSES consist of libraries and a graphical user interface supporting variety of rough set-based computations [3]. It helps in reduct generation. WEKA is used to analyze data using data mining tasks. Integrated model is generated using WEKA. Weka is used to process data.

B. *Methodology using RSES*

Reduct generation from data

In RSES Reducts can be generated using Exhaustive and Genetic algorithms. This study uses exhaustive algorithm for reduct generation which is given as follows:

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 [7]. The method is based on Boolean reasoning approach.

The algorithm is given below:

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

IV. Experimental Results

In this paper data of 4521 customers are taken from UCI repository [5] and attributes are 17 attributes. Here Rough set is used to reduce attributes. Data consist of 17 attributes out of which four attributes are reduced. For generation of reducts Exhaustive algorithm is used. The attributes reduced are 13. After reducing attributes single models are generated by using Logistic Regression, Radial Basis Neural Network, Support Vector Machine and Decision Tree [1] techniques and integrated models are generated using all these techniques. From the following tables it is observed that before reduct there is no improvement in processed integrated model accuracy with respect to unprocessed integrated model accuracy but after reduct there is an improvement in processed integrated model accuracy with respect to unprocessed integrated model accuracy. Thus with the help of reduct integrated technique improves accuracy of the model. Table II shows the comparative study of processed integrated model and unprocessed integrated model before attribute reduction while Table III shows comparative study of processed integrated model and unprocessed integrated model after attribute reduction

Table II. Comparative Study between Processed and Unprocessed Integrated Model before Reduct

	Unprocessed Data Set		Processed Data Set	
Name of Method	Single Model (Accuracy)	Integrated Model (Accuracy)	Single Model (Accuracy)	Integrated Model (Accuracy)
LR	90.24%	89.41%	90.11%	89.89%
RBF	89.27%	89.29%	89.29%	89.80%
SMO	89.16%	89.80%	89.64%	90.26%
TREE	89.49%	89.82%	91.50%	90.44%

Table III. Comparative Study between Processed and Unprocessed Integrated Model after Reduct

	Unprocessed Data Set		Processed Data Set	
Name of Method	Single Model (Accuracy)	Integrated Model (Accuracy)	Single Model (Accuracy)	Integrated Model (Accuracy)
LR	90.24%	89.69%	91.37%	91.66%
RBF	89.55%	89.29%	89.84%	91.86%
SMO	89.18%	90.02%	89.64%	91.90%
TREE	89.49%	89.82%	91.48%	90.06%

ACKNOWLEDGMENT

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

CONCLUSION AND FUTURE WORKS

Experimental results indicate that rough set technique is the best technique for attribute reduction as after reducing attributes it is found that integrated model improves accuracy. Thus it is necessary to reduce attribute which improves accuracy of the model. Future work is to generate Integrated model of data mining methods using rough set approach.

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] L. Rushi , S. Snehlata , M. Latesh : “Class Imbalance Problem in Data Mining:” Review. International Journal of Computer Science and Network (IJCSN), Vol. 2, 226-230, 2013

[3] B., Mohamed , A.,Taklit : “Imbalanced Data Learning Approaches Review.” International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol. 3, 15-33, 2013

[4] J., Kittler, M., Hatef, R. P. Duin, J., Matas : On Combining Classifiers.Pattern Analysis and Machine Intelligence, IEEE Transactions, Vol. 20, 226-239, 1998

[5] UCIrvine Machine Learning Repository

[6] Z., Pawlak: Rough sets: Theoretical aspects of reasoning about data, Kluwer Dordrecht, 1991.

[7] Rough Set Exploration System (RSES) available at http://logic.mimuw.edu.pl/ rses/.

[8] Z., Pawlak: Rough sets and decision analysis, INFOR: Information system and operational research, 38(3), 2000, 132–144.

[9] 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.

[10] Suman Saha, C. A. Murthy, Sankar K. Pal: “Rough Set Based Ensemble Classifier. Lecture Notes in Computer Science”, Vol. 6743, 27-33, 2011

[11] S., Hala , Ajith Abraham: “A New Weighted Rough Set Framework Based Classification for Egyptian Neonatal Jaundice. Applied Soft Computing.” Elsevier Vol. 12(3), 999-1005, 2012

[12] Z., Pawlak, Rough sets, Informational Journal of Computer and Information Sciences, vol.11, no.5, pp. 341-356, 1982.

[13] 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.

[14] 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.