International Journal of Database Theory and Application
Vol.7, No.5 (2014), pp.95-108
http://dx.doi.org/10.14257/ijdta.2014.7.5.07

A Rough Set Based Classification Model for the Generation of Decision Rules

Vinod Rampure[1] and Akhilesh Tiwari[2]

Department of CSE & IT, Madhav Institute of Technology and Science, Gwalior (M.P), India
[1]rampurevinod@yahoo.in, [2]atiwari.mits@gmail.com

Abstract

This paper introduces a very important classification aspect for the analysis of huge amount of data stored in databases and other repositories. Numerous classification models are available in the literature, to predict the class of objects whose class level is unknown. Literature reveals that most of the available models are not capable in handling imperfect data. In view of this, present paper proposes a new rough set based classification model to derive the classification (IF-THEN) rules. Furthermore, developed model has been applied to handle bank-loan applications database as either safe, unsafe or risky. However, proposed model can also be used for the analysis of data from other domains.

Keywords: *rough set theory, classification, reduct, core, decision rules*

1. Introduction

Data mining is the processes extracting valuable information from the huge amount of data. It is also known as knowledge discovery processes. There are various data mining task which are available in the literature, one of the important task is classification. Classification is the processes of finding a model or function that describes and distinguishes data classes, for the purpose of being able to use the model to predict the class of objects whose class level is unknown [1].

It has been observed that currently available classification models are not suitable for the imperfect data and therefore present paper proposes to make use of rough set theory. The rough set theory, introduced by pawlak in (1985) [4, 13], although popular in many discipline[5]. Rough set represent a different mathematical approach to imprecision, vagueness and uncertainty [2]. The rough set philosophy is founded on the assumption that every object of the universe of discourse. The main goals of rough set analysis are induction of approximations of concepts.

In recent years, we have witnessed a growing interest in rough set. This theory attracted attention of many researchers whose contributions have further enhanced the rough set theory. It constitutes a sound basis for decision support system and data mining. It offers solution to the problem of discretization, attribute selection, data reduction and decision rule generation. Classification approach based on rough set, called rough set based classification model which is successfully applied on real world application. This classification model performs feature selection before generating rules [9]. It is efficient technique that tries to automatically produce a minimal and significant set of decision rule without any iteration. A rough set based classification algorithm has the advantage a time complexity for learning, accuracy and size of the discovered rules.

ISSN: 2005-4270 IJDTA

This paper organized as follow: section 2 express the basic concept of rough set theory, section 3 proposed classification algorithms based on rough set. Section 4 consists of experimental analysis. Section 5 consists of result and discussion. Finally conclusion, are drawn in section 6.

2. Basic Concept of Rough Set Theory:-

A rough set methodology is based on the premise that lowering the degree of precision in the data makes the data pattern more visible [12], whereas the central premise of the rough set philosophy is that the knowledge consists in the ability of classification. In other words, the rough set approach can be considered as a formal framework for discovering facts from imperfect data [3]. The results of the rough set approach are presented in the form of classification or decision rules.

2.1. Information System

Formally, an information system IS (or an approximation space), can be seen as a system.

$$IS = (U, A)$$

Where U is the universe (a finite set of objects, U=(x1, x2……x_n) and A is the set of attributes (features, variables). Each attribute aϵ A (attribute a belonging to the considered set of attribute A) defines an information function$f_a: U \rightarrow V_a$, where V_a is the set of values of a , called the domain of attribute a.

2.2. Indiscrenibility Relation

For every set of attributes B$\subset$ A, an indiscernibility relation Ind(B) is defined in the following way: two objects, x_i and x_j, are indiscernible by the set of attributes B in A, if b(x_i) = b(x_j) for every b $\subset B$. The equivalence class of Ind (B) is called elementary set in B because it represents the smallest discernible groups of objects [8]. For any element x_i of U, the equivalence class of x_i in relation Ind(B) is represented as $[x_i]_{Ind(B)}$. The construction of elementary sets is the first in classification with rough set.

2.3. Lower and Upper Approximations

The rough sets approach to data analysis hinges on two basic concepts, namely the lower and the upper approximations of a set referring to [10]:

-the elements that doubtlessly belong to the set, and

-the elements the possibly belong to the set.

Let X denotes the subset of elements of the universe U (X $\subset$ U). The lower approximation of X in B(B$\subseteq$A) , denoted as $\underline{BX}$, is defined as the union of all these elementary sets which are contained in X.

More formally:

$$\underline{BX}=\{x_i \in U \mid [x_i]_{Ind(B)} \subset X\}$$

The above statement is to be read as: the lower approximation of the set X is a set of object x_i , which belong to the elementary sets contained in X (in the space B).

The upper approximation of the set X, denoted as BX, is the union of these elementary sets, which have a non-empty intersection with X:

BX=$\{x_i \in U \mid [x_i]_{Ind(B)} \cap X \neq 0\}$,

For any object $x_i$of the lower approximation of X (i.e,$x_i \in$BX) , it is certain that it belong to X. for any object $x_i$of the upper approximation of X(i.e,$x_i \in$BX), we can only say that $x_i$may belong to X. The difference:

BNX = BX – BX is called a boundary of X in U.

2.4. Accuracy of Approximation

An accuracy measure of the set X in B$\subseteq$A is defined as:

$\mu_B(X) = \text{card}(\underline{BX}) / \text{card}(BX)$

The cardinality of a set is the number of objects contained in the lower (upper) approximation of the set X. As one can notice, $0 \leq \mu_B(X) \leq 1$. If X is definable in U theμ_B (X) = 1, if X is undefinable in U then μ_B (X) < 1.

2.4. Core and Reduct of Attributes

In rough set theory, information table is used for describe of object in the universe, it consist of two dimensions, each row is an object, and each column is an attribute. Rough set theory classifies attribute in two types according to their roles of information table: core attribute and redundant attribute [11]. Here the minimum condition attributes set can be received, which is called reduction. One information table might have a several different reduction simultaneously. The intersection of the reduction is the core of information table and the core attribute are the important attribute that influences attribute classification [6].

A subset *B* of a set of attribute C is the reduction of C with respect to R if and only if

(1) $POS_B(R) = POS_C(R)$, and

(2) $POS_{B-\{a\}}(R) \neq POS_C(R)$, for any a$\in$B.

And the core defined by the equation given below

$$CORE_C(R) = \{c \in C \mid \forall c \in C, POS_C(R)\}$$

3. Proposed Rough set based Classification Algorithm

We propose a rough set based classification algorithm for generating classification model. The proposed algorithm is as follows:

Input: A set of n element in universe set, which have some condition attribute and one decision attribute in decision table D.

Output: generate classification model in the form of decision rules.

Step1: conditional attribute and decision attribute value provide positive integer value.

Step2: convert data positive integer form.

Step3: construction of elementary set in D-space.

Step4: calculation of upper and lower approximation of elementary sets in D.

Step5: Finding D-core and D-reduct of A attribute.

Step6: Finding D-core and D-reduct of A attribute values.

Step7: Create reduction table (Final table) of subspace reduct.

Step8: Generate classification model based decision rules.

Example:

We have used synthetic bank database, to apply classification algorithm based rough set to generate classification decision rule model. Database consists of three condition attribute age, sex and income and one decision attribute. We have build classification model to categorize a bank loan application as either safe, unsafe or risky as shown in table.

Table I. Show Bank Databases

Universe Person	Age	Sex	Income	Decision attribute (Loan Decision)
Rahul	Young	Male	High	Unsafe
Ragani	Old	Female	Low	Safe
Rastogi	Young	Male	High	Unsafe
Pooja	Young	Female	High	Safe
Ram	Child	Male	Zero	Risky
Shyam	Old	Male	Middle	Risky
Neha	Old	Female	Low	safe
Gourav	Child	Male	Zero	risky
Vinu	Young	Male	High	unsafe
Ragni	Old	Female	Low	safe

Step 1: firstly conditional attribute and decision attribute value, we assign positive integer value.

Table II. Show Table Define Integer Value, of Condition and Decision Attribute

Age	Sex	Income	Loan decision
Child=1	Male=1	Low=1	Unsafe=1
Young=2	Female=2	Middle=2	Safe=2
Old=3		High=3	Risky=3
		Zero=4	

Step 2: convert data positive integer form.

Table III. Show Table Data Positive Integer Form

Universe person	A1	A2	A3	D
1	2	1	3	1
2	3	2	1	2
3	2	1	3	1
4	2	2	3	2
5	1	1	4	3
6	1	1	2	3
7	3	2	1	2
8	1	1	4	3
9	2	1	3	1
10	3	2	1	2

Step 3: Construction of elementary set in D-space. The decision attribute d describe the belongingness of 10 objects of the following classes.

Class 1: { $x_{1,}x_{3,}x_{9}$}

Class 2: {$x_{2,}x_{4,}x_{7,}x_{10}$}

Class 3: { $x_{5,}x_{6,}x_{8}$}

Step-4: We have calculated of upper and lower approximation of elementary sets in D.

Table IV. Show Table below Upper Approximation and Lower Approximation Value and also Accuracy of Class

Class No.	No. of object	Lower approximation	Upper approximation	Accuracy
1	3	3	3	1.0
2	4	4	4	1.0
3	3	3	3	1.0

Step-5: Finding D-core and D-reduct of A attribute. One must first construct the D-discrenbility matrix, elements of which discern object from different group in D.

Table V. Show Table, we have Calculated D-core and D-reduct of Condition Attribute of A

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
X1	-------									
X2	$a_1a_2a_3$	-------								
X3	-------	$a_1a_2a_3$	--------							
X4	a_2	-------	a_2	------						
X5	a_1a_3	$a_1a_2a_3$	a_1a_3	$a_1a_2a_3$	-----					
X6	a_1a_3	$a_1a_2a_3$	a_1a_3	$a_1a_2a_3$	-----	------				
X7	$a_1a_2a_3$	-------	$a_1a_2a_3$	------	$a_1a_2a_3$	$a_1a_2a_3$	------			
X8	a_1a_3	$a_1a_2a_3$	a_1a_3	$a_1a_2a_3$	-------	------	$a_1a_2a_3$	------		
X9	--------	$a_1a_2a_3$	-------	a_2	a_1a_3	a_1a_3	$a_1a_2a_3$	a_1a_3	------	
X10	$a_1a_2a_3$	-------	$a_1a_2a_3$	-------	$a_1a_2a_3$	$a_1a_2a_3$	-------	$a_1a_2a_3$	$a_1a_2a_3$	------

The $F_A(D)$ discrinbility function has the following form:

$F_A(D)=(a_1+a_2+a_3)a_2(a_1+a_3)(a_1+a_3)(a_1+a_2+a_3)(a_1+a_3)(a_1+a_2+a_3)(a_1+a_2+a_3)(a_1+a_2+a_3)(a_1+a_2+a_3)(a_1+a_2+a_3)a_2(a_1+a_3)(a_1+a_3)(a_1+a_2+a_3)(a_1+a_3)(a_1+a_2+a_3)(a_1+a_2+a_3)(a_1+a_2+a_3)a_2(a_1+a_2+a_3)(a_1+$

$a_3)(a_1 + a_2 + a_3)(a_1 + a_2 + a_3)(a_1 + a_3)(a_1 + a_2 + a_3)(a_1 + a_2 + a_3)(a_1 + a_2 + a_3)(a_1 + a_3)(a_1 + a_2 + a_3)(a_1 + a_2 + a_3)$

Apply Absorption low

$F_A(D) = a_2(a_1 + a_3)$

$F_A(D) = a_2a_1 + a_2a_3$

There are two D-reduct=$\{a_1a_2\}$ and $\{a_2a_3\}$

Step-6: Finding D-Core and D-reduct of A attribute value ("We can also interested in elimination of unnecessary value of condition attribute in the decision table". we need to calculate relative reduct and relative core of value of attributes based on the discrniblity matrix constructed for subspace $\{a_1a_2\}$ or $\{a_2a_3\}$). Show table VI and VII.

Table VI. Show Table Reduct Subspace $\{a_1a_2\}$

U	a_1	a_2	d
X1	2	1	1
X2	3	2	2
X3	2	1	1
X4	2	2	2
X5	1	1	3
X6	1	1	3
X7	3	2	2
X8	1	1	3
X9	2	1	1
X10	3	2	2

Table VII. Show Table Reduct Subspace $\{a_2a_3\}$

U	a_2	a_3	d
X1	1	3	1
X2	2	1	2
X3	1	3	1

X4	2	3	2
X5	1	4	3
X6	1	2	3
X7	2	1	2
X8	1	4	3
X9	1	3	1
X10	2	1	2

Table VIII. Show Table, we have Elimination Unnecessary Value of Reduct $\{a_1a_2\}$

	1	2	3	4	5	6	7	8	9	10
1	------	a_1a_2	----	a_2	a_1	a_1	a_1a_2	a_1	----	a_1a_2
2	a_1a_2	------	a_1a_2	--------	a_1a_2	a_1a_2	-----	a_1a_2	a_1a_2	------
3		a_1a_2	----	a_2	a_1	a_1	a_1a_2	a_1	----	a_1a_2
4	a_2	a_1	a_2	-------	a_1a_2	a_1a_2	------	a_1a_2	a_2	-----
5	a_1	a_1a_2	a_1	a_1a_2	-----	-------	a_1a_2	------	a_1	a_1a_2
6	a_1	a_1a_2	a_1	a_1a_2	------	------	a_1a_2	--------	a_1	a_1a_2
7	a_1a_2	------	a_1a_2	------	a_1a_2	a_1a_2	------	a_1a_2	a_1a_2	------
8	a_1	a_1a_2	a_1	a_1a_2	------	------	a_1a_2	-----	a_1	a_1a_2
9	------	a_1a_2	------	a_2	a_1	a_1	a_1a_2	a_1	-----	a_1a_2
10	a_1a_2	------	a_1a_2	------	a_1a_2	a_1a_2	------	a_1a_2	a_1a_2	----

$F_1(D)= (a_1 + a_2)a_2a_1a_1(a_1 + a_2)a_1(a_1 + a_2)=a_1a_2$

$F_2(D)= (a_1 + a_2)(a_1 + a_2)a_1(a_1 + a_2)(a_1 + a_2)(a_1 + a_2)(a_1 + a_2)=(a_1+a_2)$

$F_3(D)=(a_1 + a_2)a_2a_1a_1(a_1 + a_2)a_1(a_1 + a_2)=a_1a_2$

$F_4(D)=a_2a_2(a_1 + a_2)(a_1 + a_2)(a_1 + a_2)a_2=a_2$

$F_5(D)=a_1(a_1 + a_2)a_1(a_1 + a_2)(a_1 + a_2)a_1(a_1 + a_2)=a_1$

$F_6(D)= a_1(a_1 + a_2)a_1(a_1 + a_2)(a_1 + a_2)a_1(a_1 + a_2)=a_1$

$F_7(D)=(a_1+a_2)(a_1+a_2)(a_1+a_2)(a_1+a_2)(a_1+a_2)(a_1+a_2)=a_1+a_2$

$F_8(D)=a_1(a_1+a_2)a_1(a_1+a_2)(a_1+a_2)a_1(a_1+a_2)=a_1$

$F_9(D)=(a_1+a_2)a_2a_1a_1(a_1+a_2)a_1(a_1+a_2)=a_1a_2$

$F_{10}(D)=(a_1+a_2)(a_1+a_2)(a_1+a_2)(a_1+a_2)(a_1+a_2)(a_1+a_2)=a_1+a_2$

Step7: Create reduction table (Final table) of subspace reduct $\{a_1a_2\}$.

Table IX. Show the Final Table of Reduct $\{a_1a_2\}$

U	a_1	a_2	D
X1	2	1	1
X2	*	2	2
X3	2	1	1
X4	*	2	2
X5	1	*	3
X6	1	*	3
X7	*	2	2
X8	1	*	3
X9	2	1	1
X10	*	2	2

Step8: Generate classification model based decision rules

$$a_{1_2}a_{2_1}=d_1$$

$$a_{2_2}=d_2$$

$$a_{1_1}=d_3$$

1. If unsafe loan decision then age is young and sex is male.
2. If safe loan decision then sex is female.
3. If risky loan decision then age is child.

4. Experimental Analysis:

This section describes the experimental outcome of the developed classification model based on rough set. We use ROSE2 system is a successor of RoughDAS and

Rough Class systems. RoughDAS is historically one of the first successful implementations of the rough set theory, which has been used in many real life applications.

The system contains several tools for rough set based knowledge discovery, *e.g.*:

- data preprocessing, including discretization of numerical attributes,
- performing a standard and an extended rough set based analysis of data,
- search of a core and reducts of attributes permitting data reduction,
- inducing sets of decision rules from rough approximations of decision classes,
- evaluating sets of rules in classification experiments,
- Using sets of decision rules as classifiers.

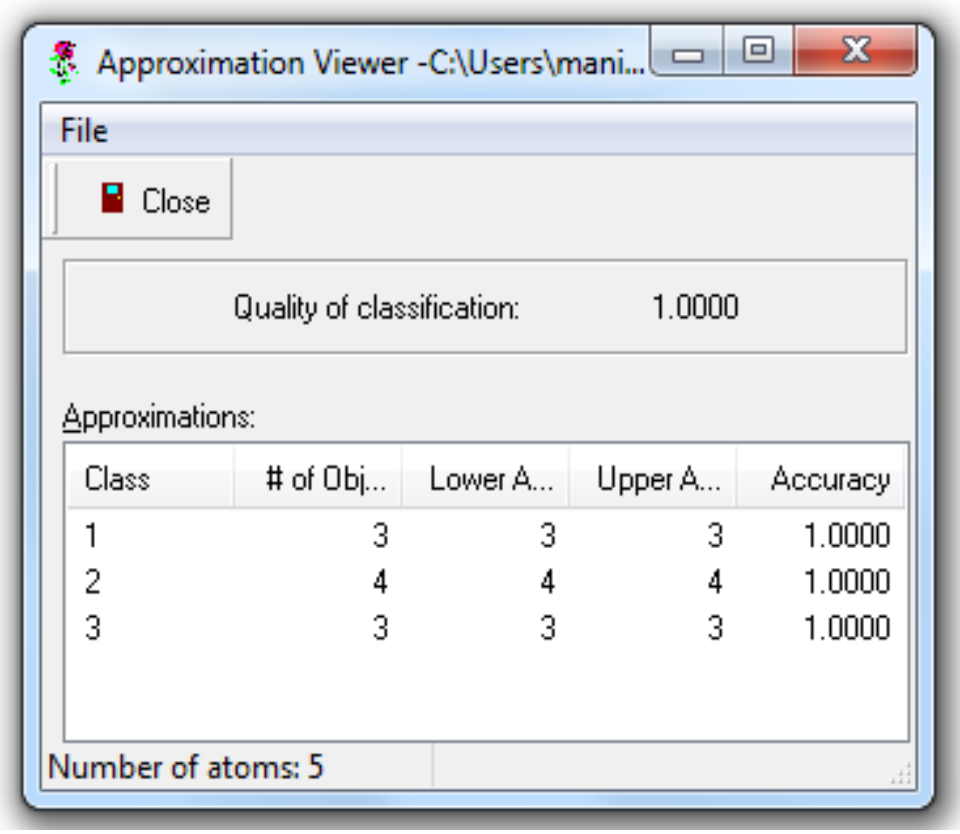

Figure 1. Shows the Lower Approximation, upper Approximation and Classification Accuracy

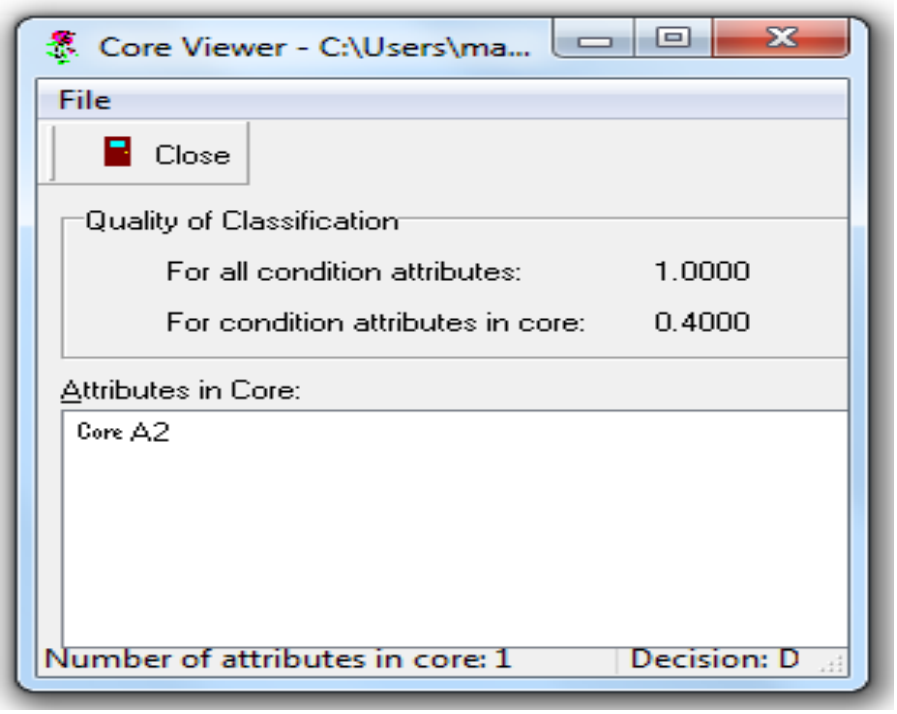

Figure 2. Shows the Number of Reduct by using Discernibility Matrix

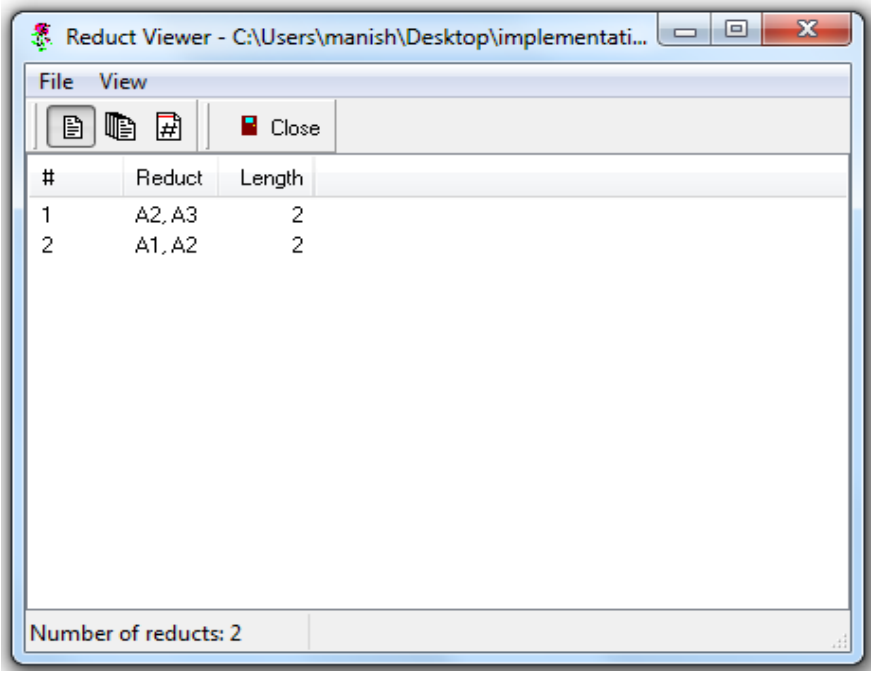

Figure 3. Show the Quality of Classification for all Condition Attributes and for Condition Attributes in Core A2

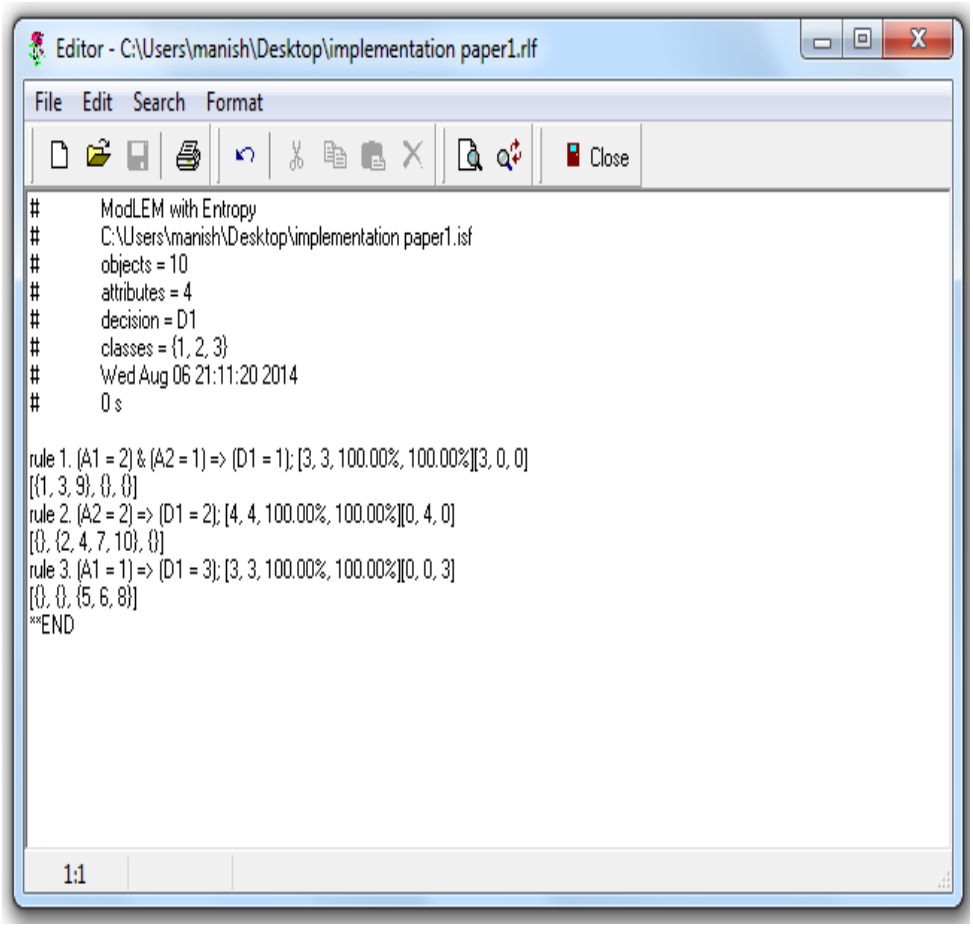

Figure 4. Show the Classification Model which Generate Decision Rules

5. Conclusion

In this paper, we have proposed a new classification model which is based on rough set. Classification algorithm, usually called supervised learning in data mining, is one of the most important and fundamental function in data analysis. The ultimate goal of classification is to build a set of models that can be used for the predict class of different object. Present work proposes a new rough set based classification model for handling imperfect (imprecision, vague, incomplete) data. Experimentation has been performing considering bank loan application database. Experimental results include generated (If-Then) rules which are useful in important decision making task. The key principal of classification is to extract classification model of training data set, and then classify unknown data using this model. The proposed algorithm can widely used in data analysis and other application domain. A concrete instance verified the feasibility.

References

[1] L. A. Zadeh, fuzzy sets, inf. Control 8 **(1965)**, pp. 338-353.
[2] B. Walczak, D. L. Massart, "tutorial: rough set theory", Chemometrics and intelligent laboratory system, vol. 47, **(1999),** pp. 1-16.
[3] J. R. Quinlan, J.Ross, in: R.S. Michalski, *et al*. (Eds.), "Machine Learning: An Artificial Intelligence Approach", Tioga, Palo Alto, **(1983)**.
[4] R. Slowinski (Ed.), "Intelligent Decision Support", Handbook of Application and Advances of the rough set theory, Kluwer Academic Publishers, Dorderecht, **(1992).**
[5] Z. Pawlak, "Rough set", Int.j.Inf. Comput. Sci, vol. 11, **(1982),** pp. 341-356.
[6] W. P Ziarko (Ed.), Rough Sets, Fuzzy sets and Knowledge Discovery, Springer, New York, **(1994)**.
[7] Z. Pawlak, Rough Sets, Theoretical Aspects of Reasoning about Data, Kluwer Academic Publisher, Dordrecht, Netherlands, **(1991).**
[8] T. Jian-guo, T. Ming-shu, "On finding core and reduction in rough set theory", Control and Decision, vol. 18, no.4, **(2003)**, pp. 449–457.
[9] Di Kai-chang Li De-ren and Li De-yi, "Rough set theory and its application in attribute analysis and kenowledge discovery in gis", Journal of wuhan technical university of surveying and mapping, vol. 24, no.1, **(1990)**, pp. 1–10.
[10] S. Xiao-xue, "Rough sets theory and its application," journal of xianyang normal university, vol. 20, no.2, **(2005)**, pp. 30–35.
[11] S Kotsiantis, D. Kanellopoulos and P.Pintelas, "Data preprocessing for supervised leaning," Internal Journal of Computer Science, **(2006)**, vol.1, no.2.
[12] Y. Yao and Y. Zhao, "Attribute reduction in decision-theoretic rough set models", Information sciences, vol.178, no. 17, **(2008)**, pp. 3356-3373.
[13] Y. Yao, "Notes on rough set approximations and associated measures", Journal of zhejiang ocean university(natural science), vol. 29, no. 5,**(2010)**, pp. 399-410.
[14] L.-Xiang-wei,Q.Yian-fang " A Data Preprocessing algorithm for classification Model Based on rough sets", International conference on Solid state Device and Material Science **(2012)**.

Authors

Vinod Rampure is currently pursuing the M.tech degree at the department of CS/IT from MITS Gwalior (M.P), India. He received him B.tech degree from Mahatma Gandhi chitrakoot university, chitrakoot satna (M.P), India. Him current interest in data mining, rough set, classification and their application.

Dr. Akhlesh Tiwari has received Ph.D. degree in Information Technology from Rajiv Gandhi Technological University, Bhopal, M.P. (India). He is currently working as Associate Professor in the Department of CSE & IT, Madhav Institute of Technology & Science (MITS), Gwalior, India. He has guided several theses at Master and Under Graduate level. His area of current research includes Knowledge Discovery in Databases and Data Mining, Wireless Networks. He has published more than 20 research papers in the journals and conferences of international repute. He is also acting as a reviewer & member in editorial board of various international journals. He is having the memberships of various Academic/ Scientific societies including IETE, CSI, GAMS, IACSIT and IAENG.