Reason - A Technical Journal
ISSN 2277–1654
Volume - XIV ● 2015

ROUGH SET: A CASE STUDY ON TECHNICAL TEACHERS' TRAINING DATA

Sukanta Ghosal[1] and Samir Roy[2]

National Institute of Technical Teachers' Training & Research, Kolkata.
Email: [1]sukanta.ghosal@outlook.com, [2]samir.cst@gmail.com

Paper received on: December 12, 2015, accepted after revision on: March 18, 2016

Abstract: This paper presents the theory of rough sets as a mathematical model of vague or inexact data. As well as it's application as a data mining tool for rule generation, out of large data. The domain of data has been taken from the training data of National Institute of Technical Teachers' Training & Research, Kolkata. ROSETTA, a rough set theoretic application tool developed at Warsaw University, Poland, has been used for the purpose. The rules generated through experiment reveal interesting knowledge hidden in the un-organized data of the target domain.

Keywords : Indiscernibility; Boundary Regions; Reducts; ROSETTA; Rule Extraction.

1. INTRODUCTION

Rough set theory (RST) proposed by Zdzislaw Pawlak in 1982, is a mathematical tool to handle imperfect Data. This method has been developed to manage uncertainties from information that represents some in exactitude, incompleteness, and vagueness. In RST vagueness or inexactness is expressed in terms of boundary regions of a set of objects. There are some situation related to large multidimensional data, when it is not possible to decide with certainty whether a given object belongs to a set or not. These objects is said to form a boundary region for the set. If the boundary region is empty, then the set is crisp. Otherwise it is rough.

There are various application of RST in field of AI, like machine learning, Data mining, pattern recognition etc. Some of the applications of RST, used in data mining will be discussed in this paper. Data mining is the process of discovering insightful, interesting, and novel patterns, as well as descriptive, understandable, and predictive models from large-scale data.

2. BASICS OF ROUGH SET THEORY

2.1 Information System

'Information' is processed or structured data that may convey some meaning. Let A = (A_1, A_2, A_3, … , A_k) be a non-empty finite set of attributes and U = {(a_1, a_2, ..., a_k)} be a non-empty finite set of k-tuples, termed as the objects. V (A_i) denote the set of values for the attributes A_i. Then an **information system** is defined as an ordered pair I (U, A) such that for all i = 1, 2, ..., k there is a function f_i.

Table 1: An Information System: Employee Expenditure Data

#	Name	Salary	Bank Savings	Food Expenditure	Luxury Expenditure	Property Values	Insurance
1	SMITH	High	High	High	Average	High	Yes
2	JONES	High	High	Average	Average	High	No
3	DAVIS	Low	Low	High	Low	High	Yes
4	CLARK	Low	High	Average	Low	Low	No
5	CARTER	High	Low	High	Average	High	Yes

f_i : U ’& V(A_i)

Table 1: shows an information system regarding the various expenditure and property of some employee of an Organization. The information system consists of five objects, each corresponding to an employee. Here U includes the objects (SMITH, High, High, High, Average, High, Yes), .., (CARTER, High, Low, High, Average, High,Yes) and the set A consists seven attributes viz., Name, Salary, Bank Savings, Food Expenditure, Luxury Expenditure, Property Value, Insurance. V (Name) = {SMITH, JONES, DAVIS, CLARK, CARTER} , V (Salary) = V (Bank Savings) = V(Luxury Expenditure) = V(Property Value)= {High, Average, Low} and V(Insurance)= {Yes, No}.

2.2 Blocks of attribute value pairs

Let one consider a attribute-value pair t = (a, v) where $a \in A$, $v \in V$. The block (denoted by [t]) denotes the set of all cases from where each attribute ‘a’ has a value ‘v’. In the association rule approach of data mining, the support measure of an attribute, compute the existence of an attribute in a specified row, then the support of an attribute-value pair is obtained by the cardinality of [τ] (|[τ]|). In the information table shown in Table 1, the block and support are defined as follows:

[(Salary, High)] = {1,2,5}, and support ([(Salary, High)])=3

[(Salary, Low)] = {3,4}, and support([(Salary, Low)])=2

[(Bank Savings, High)] = {1,2,4}, and

support ([(Bank Savings, High)]) = 3, and so on.

2.3 Decision system

Decision System has the capacity to take decision from an information system. A Decision System D(U,A,d) is an information system I(U,A) augmented with a special attribute d $\notin$ A, known as the Decision attribute. In the table 2, a special attribute ‘Willing to Buy Car’ is added with table 1, here the attribute ‘Willing to Buy Car’ is a decision attribute.

2.4 Indiscernibility

Let I = (U, A) be an information system where U = {(a_1, …, a_k)} is the non-empty finite set of k-tuples known as the objects and U = {A_1, …, A_k} is a non-empty finite set of attributes. Let P $\subseteq$ A be a ∅ subset of the attributes. Then the set of P-indiscernible objects is defined as the set of objects having the same set of attribute values.

IND_I (P) = {(*x*, *y*), *x*, *y* $\in$ U | $\forall$ a $\in$ A, *x* (a) = *y*(a)}

Consider the Table 1, Let P= {Food Expenditure, Property Value, Insurance} belongs to A={ Salary, Bank Savings, Food Expenditure, Luxury Expenditure, Property Value, Insurance }. IND_I (P)={(SMITH, DAVIS, CARTER),(JONES),(CLARK)}.

Table 2: A Decision System: Employee Expenditure Data

#	Name	Salary	Bank Savings	Food Expenditure	Luxury Expenditure	Property Value	Insurance	Willing to Buy Car
1	SMITH	High	High	High	Average	High	Yes	Yes
2	JONES	High	High	Average	Average	High	No	Yes
3	DAVIS	Low	Low	High	Low	High	Yes	No
4	CLARK	Low	High	Average	Low	Low	No	No
5	CARTER	High	Low	High	Average	High	Yes	Yes

2.5 Approximations

In a decision system, the indiscernibility equivalence relation partitions the universe U into a number or subsets based on identical values of the outcome attribute. Such partitions are crisp and have clear boundaries. However, such crisp partitions might not be always possible. For example, consider the decision system presented in Table 3. It consists of Prices of some Cars ranging between 4 to 12 Lac. The outcome attribute 'Air Bag for all Seat' has the possible values of YES or NO depending on whether all the seats of the car having Air Bag or not.

Table 3: Car Feature Information

#	Car Price (In Lac)	Air Bag for All Seats
1	4	No
2	9	Yes
3	7	No
4	12	Yes
5	4	No
6	9	Yes
7	12	Yes
8	7	Yes

Consider the entries 3 & 8, where the price are same but the outcome is different. Under Such circumstances the concept of rough sets comes into the picture. Rough sets are defined in term of lower and upper approximations. These are described below:

2.5.1 Lower and Upper Approximations

Let I = (U, A) be an information system and B $\subseteq$ A and X $\subseteq$ U. Then

B-lower approximation of X= $\underline{B}$ (X) ={x | $[x]_B$ $\subseteq$ X }

B-upper approximation of X= $\overline{B}$(X)={x| [x]B)"X ‘“ $\varnothing$}

2.5.2 Boundary Region

The set BN_B (X) = $\overline{B}$X − $\underline{B}$X is called the B-boundary region of X. The B-boundary region of X consists of those objects which one cannot decisively classify as inside or outside the set X on the basis of the knowledge of their values of attributes in B. If a set has a non-empty boundary region, it is said to be a rough set.

2.5.3 Outside Region

The set U − $\overline{B}$X is called the B-outside region of X. The B-outside region of X consists of elements that are classified with certainty as not belonging to X on the basis of knowledge in B.

With reference to the information system presented in Table 3, let W = {y | Air Bag for All Seats (y) = Yes} ={2, 4, 6, 7, 8}. Now, the set of Price-indiscernible objects of U, IND_{Price} (U) = {{1, 5}, {2, 6}, {3, 8}, {4, 7}}. Hence the sets of the Price-indiscernible objects for various objects are $[1]_{Price}$ = $[5]_{Price}$ = {1, 5}, $[2]_{Price}$ =$[6]_{Price}$ = {2, 6}, $[3]_{Price}$ = $[8]_{Price}$ = {3, 8}, $[4]_{Price}$ = $[7]_{Price}$ = {4,7}. Thus, assuming B = {Price} one has

B-lower approximation of W : $\underline{B}$W = {2, 4, 6, 7}

B-upper approximation of W : $\overline{B}$W = {2, 3, 4, 6, 7, 8}

B-boundary region of W : BN_B (W) = {3, 8}

B-outside region of W : U − $\overline{B}$W = {1, 5}

3. APPLICATIONS

So far in this paper, some of the theory aspects of RST has been discussed, which can be applied in Data Mining. Now, some real time application of RST in data mining will be shown in this section, i.e finding minimal Reduct from an information system, rule generation.

3.1 Reducts

A minimal set of attributes required to preserve the indiscernibility relation among the objects of an information system is called a reduct. Given an information system I = (U, A), a reduct is a minimal set of attributes B $\subseteq$ A such that IND_I (B) = IND_I (A). A Reduct with minimal cardinality is called a minimal reduct.

Consider the Decision System in Table 2. (Excluding the decision attribute ''Willing to Buy Car") .The minimal reduct has been developed from the decision system, and shown in Table 4. The decision system have 5 objects, so the Discernibility Matrix will be a 5 X 5 matrix, where the (i,j)th element d_{ij} is given by $d_{ij}=\{a \in A|\ a(x_i) \neq a\ (x_j)\}$. Each entry of a discernibility matrix is one or more attributes for which the objects x_i and x_j differ.

***Note:** Here in the Discernibility Matrix, the Characters (S, B, F, L, P, I) represents the attributes Salary, Bank Savings, Food Expenditure, Luxury Expenditure, Property Value, Insurance.*

The discernibility function for the discernibility matrix shown in Table 4, will be generated now. A discernibility function f_I for an information system I=(U,A) is a Boolean variables $a_1,a_2,...a_n$ corresponding to the *n* number of attributes $A_1,......A_n$ such that

$$f_I\ (a_1,a_2,...a_n)= \{\ V\ d_{ij} \mid 1 \leq i\ n,\ d_{ij} \neq \Phi\ \}$$

Where d_{ij} is the (i, j)th entry of the discernibility matrix. The set of all prime implicants corresponds to the set of all reducts of I. Hence, the aim is to find the prime implicants of f_I. The discernibility function for the discernibility matrix shown in Table 4 is given in Figure 1.

The prime implecants are $(B \wedge F \wedge S)$, $(B \wedge F \wedge L)$, $(B \wedge I\ v\ S)$ and $(B \wedge I \wedge L)$

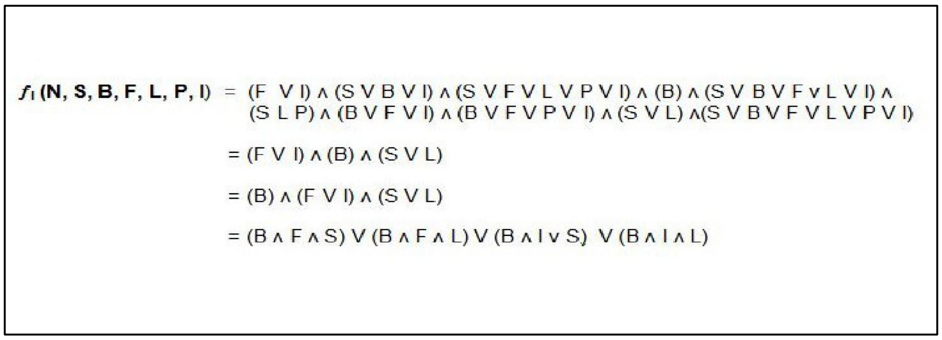
Each of the sets {B, F, S}, {B, F, L}, {B, I, S}, and {B, I, L} are the minimal set of attributes that preserves the classification IND_I (A). Hence each of them is a reduct. As all the reducts are of size 3, hence all the reducts are minimal reducts.

3.2 Rule Generation

Let one consider the decision system D=(U,{Name, Salary, Bank Savings, Food Expenditure, Luxury Expenditure, Property Value, Insurance}, {Willing to Buy Car}) concerning Employee Expenditure Information, as shown in the table 2. The minimal reducts for the same system has been calculated in the section 2.3.The sample space after attribute reduction (using the minimal reduct set {Bank Savings, Food Expenditure, Salary}, however any of the minimal reduct set can be used) is shown in the Table 5, The extracted rules using the minimal reduct obtained above are given in the Table 6

Table 4: Disceribility Matrix for Employee Expenditure Data

	1	2	3	4	5
1	∅	F, I	S, B, L	S, F, L, P, I	B
2	-	∅	S, B, F, L, I	S, L, P	B, F, I
3	-	-	∅	B, F, P,I	S, L
4	-	-	-	∅	S, B, F, L, P, I



$$\begin{aligned} f_I\ (N, S, B, F, L, P, I) = {} & (F \vee I) \wedge (S \vee B \vee I) \wedge (S \vee F \vee L \vee P \vee I) \wedge (B) \wedge (S \vee B \vee F\ v\ L \vee I) \wedge \\ & (S\ L\ P) \wedge (B \vee F \vee I) \wedge (B \vee F \vee P \vee I) \wedge (S \vee L) \wedge (S \vee B \vee F \vee L \vee P \vee I) \\ = {} & (F \vee I) \wedge (B) \wedge (S \vee L) \\ = {} & (B) \wedge (F \vee I) \wedge (S \vee L) \\ = {} & (B \wedge F \wedge S) \vee (B \wedge F \wedge L) \vee (B \wedge I\ v\ S)\ \vee (B \wedge I \wedge L) \end{aligned}$$

Fig. 1: The Discernibility function

4. RULE EXTRACTION: A CASE STUDY ON NITTTR, KOLKATA TRAINING DATABASE

One of the most popular application of data-mining is Rule extraction from an large database. So far in this paper some of the basics of rough set theory has been discussed. Now later in this section, how the rough set theory will help to extract rules from a large database will be discussed.

4.1 Problem Domain

National Institute of Technical Teachers' Training & Research runs a short term training program in the Institute premises & at two of its' extension centre located at Bhubaneswar & Guwahati . Teachers from different government & private colleges from different part of the country attend these training programs. Each training program is conducted by the faculty members of the institute. The institute offers more than 100 training programs on an average per academic year. So naturally thousands of teachers(trainee) took participation in those short term training programs. Various data about these trainees can be collected and an information system can be constructed from that. Then some knowledge can be extracted from the information system in term of rules.

Let one take some training details for some of the trainee who have attended the STTP, and constitute two Decision systems with two different decision attributes.

Let one consider an Decision system (shown in Table 7) I = (U, { Designation, Department, State, Course, Course Co-ordinator} { Belongs to North Indian State}), describing training details of trainees. Here in this example, some of the attributes has been considered from the trainee database, just to illustrate the methodology. Here in this example, the authors have shown the Designation of the trainee, the Department and the State of the trainee, the Course he/she has enrolled, the Course Co-ordinator and a Decision attribute 'Belongs to North Indian State' describing whether the Trainee belongs to a North Indian State or not. In the real system, there are few more attributes like Course duration, the Institute Type(Govt/Pvt) from where the trainee belongs to, Contact No, email etc. The real system consist of approximately 4000 records for about 103 courses. Here in this example a sub part has been taken from the database, describing the details of five trainee.

Now, another decision attribute 'Training Duration 1 Week' has been taken and constructed a decision system shown in Table 8.

Table 5: Sample Space After Attribute Reduction

#	Bank Savings	Food Expenditure	Salary	Willing to Buy Car
1	High	High	High	Yes
2	High	Average	High	Yes
3	Low	High	Low	No
4	High	Average	Low	No
5	Low	High	High	Yes

Table 6 : Extracted Rules

Rule #	**Antecedents**	**Consequent**
1	**IF** (Bank Saving=High), (Food Expenditure = High/Average) & (Salary = High)	Willing to Buy Car= Yes
2	**IF** (Bank Savings= Low), (Food Expenditure = High) & (Salary= Low)	Willing to Buy Car= No
3	**IF** (Bank Savings= High), (Food Expenditure = Average) & (Salary= Low)	Willing to Buy Car= No
4	**IF** (Bank Savings= Low), (Food Expenditure = High) & (Salary= High)	Willing to Buy Car= No

Table 7: Decision System: Trainee Information

#	Name	Designation	Department	State	Course	Course Co-ordinator	Belongs to North Indian State
1	Brajendra Kumar	Lecturer	CSE	Jharkhand	Topics on Artificial Intelligence	Dr. S. Roy	No
2	S. H. Ahmed Azad	Lecturer	CSE	Assam	Design of Business application using S.E approach	Dr. R. Dasgupta	Yes
3	Sanchita Ghosh	Assistant Professor	CSE	W.B.	Introduction to DBMS	Dr. R. Dasgupta	No
4	N. Renuka Devi	Lecturer	Architecture	Manipur	Topics on Artificial Intelligence	Dr. S. Roy	Yes
5	Amlan R. Choudhury	Assistant Professor	CSE	Tripura	Introduction to Soft Computing	Dr. S. Roy	Yes

Table 8: Decision System: Trainee Information

#	Name	Designation	Depart-ment	State	Course	Course Co-ordinator	Training Duration 1week
1	Brajendra Kumar	Lecturer	CSE	Jhar-khand	Topics on Artificial Intelligence	Dr. S. Roy	Yes
2	S. H. Ahmed Azad	Lecturer	CSE	Assam	Design of Business application using S.E. approach	Dr. R. Dasgupta	No
3	Sanchita Ghosh	Assistant Professor	CSE	W.B.	Introduction to DBMS	Dr. R. Dasgupta	No
4	N. Renuka Devi	Lecturer	Archite-cture	Manipur	Topics on Artificial Intelligence	Dr. S. Roy	Yes
5	Amlan R. Choudhury	Assistant Professor	CSE	Tripura	Introduction to Soft Computing	Dr. S. Roy	No

The Decision system (shown in Table 8) is constructed with the decision attribute 'Training Duration 1week' indicates whether the duration of the training is of 1week or not.

Let's try to find some knowledge In-term of Rule Extraction based on the information system. As far in this article the basic property of rough set has been discussed, and the rule extraction method based on Minimal Reduct. So in order to extract rule from the database one needs to calculate the Minimal Reduct set.

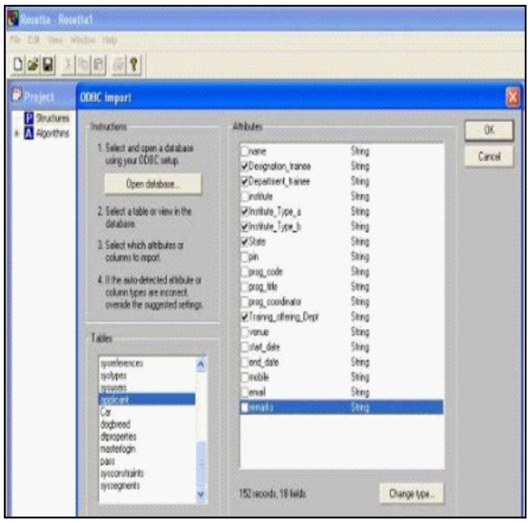
4.2 ROSETTA : Rough Set Software System

Research in Rough sets has resulted in a number of software tools for data mining and knowledge discovery from database (KDD). Among many of these tools the **ROSETTA** system is probably one of the most complete software environment for rough set operations.

The ROSETTA system has been developed by two groups: Knowledge System Group at NTNU, Norway, and Group pf Logic, Warsaw University, Poland, under the guidance of, respectively, Jan Komorowski and Andrzej Skowron.

In **ROSETTA**, the experimental nature of rough set including classifiers form data is explicitly maintained by organizing the workspace in a tree structure that display how input and output data relate to each other. ROSETTA supports the overall KDD process: From browsing and preprocessing of the data, to reduct computation and rule synthesis, to validation and analysis of the generated rules.

To extract the rules from the NITTTR, Kolkata Training database ROSETTA can be used. Among several algorithms for calculating Reducts in ROSETTA system, *RSESExhaustive Reducer (Exhaustive Calculation)* is being used to extract the rules from the database.



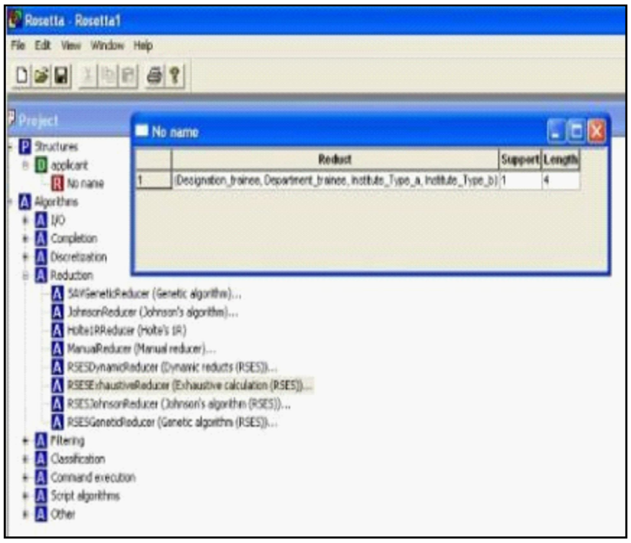
1. Select and open a database using your ODBC setup.
2. Select a table or view in the database.
3. Select which attributes or columns to import.
4. If the auto-detected attribute or column types are incorrect, override the suggested settings.

Tables: sysreferences, systypes, sysusers, applicant, Car, dogbreed, dtproperties, masterlogin, pass, sysconstraints, syssegments

□ name	String
☑ Designation_trainee	String
☑ Department_trainee	String
□ institute	String
☑ Institute_Type_a	String
☑ Institute_Type_b	String
☑ State	String
□ pin	String
□ prog_code	String
□ prog_title	String
□ prog_coordinator	String
☑ Training_offering_Dept	String
□ venue	String
□ start_date	String
□ end_date	String
□ mobile	String
□ email	String
□ remarks	String

152 records, 18 fields

Fig. 2: Choosing attribute in ROSETTA


Rosetta - Rosetta1
File Edit View Window Help
Project
No name
Structures
applicant
No name
Algorithms
I/O
Completion
Discretization
Reduction
SAVGeneticReducer (Genetic algorithm)...
JohnsonReducer (Johnson's algorithm)...
Holte1RReducer (Holte's 1R)
ManualReducer (Manual reducer)...
RSESDynamicReducer (Dynamic reducts (RSES))...
RSESExhaustiveReducer (Exhaustive calculation (RSES))...
RSESJohnsonReducer (Johnson's algorithm (RSES))...
RSESGeneticReducer (Genetic algorithm (RSES))...
Filtering
Classification
Command execution
Script algorithms
Other

	Reduct	Support	Length
1	{Designation_trainee, Department_trainee, Institute_Type_a, Institute_Type_b}	1	4

Fig. 3: Minimal Reduct set in ROSETTA

In the case study on NITTTR,Kolkata database, ROSETTA has been used to calculate the minimal reduct set. Here the authors have taken 170 objects from the database with 18 attributes, among which authors have taken 6 attributes namely {Designation_trainee, Department_trainee, State, Training_offering _dept, Institute_type_a, Institute_type_b} [Fig 2].

By using RSES Exhaustive reducer algorithm in ROSETTA authors have found a minimal reduct set of length four i.e {Designation_ trainne, Department_trainee,Institute_type_a, Institute_type_b} [Fig 3].

5. CONCLUSION

Fundamentals of Rough Set Theory and it's application for data mining has been discussed in this paper. The standard application tool for rough set namely ROSETTA, has been explored. The proposed methodology for rule extraction has been exploited for extraction of knowledge from large training data. Interesting hidden knowledge have been revealed through the application.

REFERENCES

[1] Roy, S. and Chakraborty, U., Introduction to Soft Computing: Neuro-Fuzzy and Genetic Algorithms, Pearson, 2013.

[2] Slimani T., Application of Rough Set Theory in Data Mining, International Journal of Computer Science & Network Solutions, Vol. 1, No.3, pp.1-10, 2013.

[3] Pawlak, Z., Rough Sets, International Journal of Computer & Information Sciences, Vol. 11, No.5, pp.341-356, 1982.

[4] Grzymala-Busse, J.W., Rough Set Theory with Applications to Data Mining, Chapter-Real World Applications of Computational Intelligence, Vol. 179, Studies in Fuzziness and Soft Computing, pp.221-244, 2005.

[5] Komorowski, J., ∅hrn, A. and Skowron, A., The ROSETTA Rough Set Software System.

[6] Hvidsten T.R., A Tutorial-Based Guide to the ROSETTA System: A Rough Set Toolkit for Analysis of Data, May 2006.