CORE GENERATION FROM PHONE CALLS DATA USING ROUGH SET THEORY

Sanjiban Sekhar Roy [1], **Sarvesh S S Rawat** [2]

[1] **School of Computing Science and Engineering, VIT University,Vellore, India**
[2] **School of Electronics and Instrumentation, VIT University,Vellore, India**

***ABSTRACT*:** During the year 1982, rough set theory was introduced by Z Pawlak in order to deal with uncertain data. Rough set theory relies only on the available data and attributes to analyze features as well as to generate classification rules without any additional information. Here we have used rough set theory to find out reduct and core of the call center data by reducing the superfluous information, and then provided an algorithm based upon the hidden pattern in data which will adequately increase the efficiency of the organization.

KEYWORDS: Rough set, Uncertain data, Reduct and Core.

1. INTRODUCTION

Call center is an organization which is responsible for delivering and transmitting huge amount of telephone calls. It is controlled by another organization to manage product prop up from consumers. This paper's main aim is to analyze the call center's information to reduce the redundant of data and to discover the abnormal patterns. Rough sets theory is known to reduce the number of attribute values without affecting the original results. Data from a call center has been used [C+01] in this study and a systematic process in the ambit of rough set theory been applied to produce the final decision rules as it helps in generating minimal set of rules from the raw data that is quite easy to understand. Besides, Rough Set Theory reduces the needed number of attribute values to produce a more compact decision rule set and increases efficiency.

2. A SKETCH OF ROUGH SET

The theory of Rough set is a new mathematical tool to deal with intelligent data analysis and data mining [Paw01] proposed by Z Pawlak. This theoretical framework is based on the concept that every object in the universe is attached with some kind of information. Set theory is a great help to the computer science research and theory of Rough set is an extension to that. It's a mathematical tool to deal with inexact, uncertain or vague data, which are part of artificial intelligent system. It includes algorithms for generation of rules, classification and reduction of attributes. It is hugely used for knowledge discovery [SD01] and reduction of knowledge. The theory of Rough set has got many important applications.

3. APPROXIMATION SPACES FOR ROUGH SET

The definition of approximation spaces and basic mathematics related with rough set are given as follows -

Definitions [Paw01]:

Let us say that the universe of discourse is U and also assume that R is equivalence relation based on U. So the approximation space which is a pair<U,R>, where U={x_1,x_2,…x_i,…..,x_n}, where x is an element such that x ε U. We consider two subsets namely lower and upper approximations which can be defined as

$$\underline{R}X = \{x \mid [x] \subseteq X\}$$

$$\bar{R}X = \{x \mid [x] \cap X \neq \phi\}$$

The pair RX= ($\underline{R}X, \bar{R}X$) is known as rough set of X. Also sometime $\underline{R}$X and $\bar{R}$X are called R-lower and R-upper approximation spaces of X. Again the sets

POS$_R$= $\underline{R}X$, NEG$_R$=U-$\bar{R}X$ and BND$_R$= $\underline{R}X$ - $\bar{R}X$

can be introduced as R-positive, R-negative and R-boundary region of X respectively.

4. INDISCERNIBILITY

Indiscernibility is an equivalence relation. A binary relation $R \subseteq X \times X$ is reflexive (*xRx* for any object *x*) is symmetric (if *xRy* then *yRx*), and is transitive (if *xRy* and *yRz* then *xRz*). The equivalence class

$[x]_R$ of an element consists of all objects $y \in X$ such that *xRy*.

Let *IS = (U, A)* be an information system, then with any $B \subseteq A$,

$$IND_{IS}(B) = \{(x,x') \in U^2 \mid \forall a \in B, a(x) = a(x')\}$$

where, $IND_{IS}(B)$ is called the *B-indiscernibly and* $(x,x') \in IND_{IS}(B)$. If $(x,x') \in IND_{IS}(B)$ then objects *x* and *x'* are indiscernible from each other by attributes from *B*. The equivalence classes of the *B-indiscernibility relation* are denoted by $[x]_B$.

5. ANALYSIS OF DATA FROM CALL CENTRE BY APPLYING MATHEMATICAL TOOLS OF ROUGH SET

A call centre handles huge amount of data. In a call centre, an agent after serving and solving the problem of each caller makes the entry of the data received, in the form of a customer service ticket, and issue separate new customer service ticket for every new problem. A summary of the data is generated to record the success and failure of this event.

Example:

Table 1 consists of 11 service ticket summaries. The columns and rows in this table are attributes and objects, respectively, and Ticket ID is a sequentially given number that is used as the main attribute. Agent ID, Product ID, Complaint type, and Service type that are shown here are condition attributes. Every row consists of a service ticket summary. These service ticket summaries contain many attributes. Their values are taken from the responses which have been recorded in service tickets. Here we have used only a few portion of attributes, and these attributes are following:

Agent ID - Agents ID is divided into temporary (T) and permanent (P). For example, T01 Means the temporary work – study students. Product ID –The company is a service provider for the business products. The company offers 36 business products, and the product that is creating problems to customer is reported here. Complaint type – the issues of customers are classified by agents. Service type –it contains the types of calls, called as inbound (i.e, customer called) and outbound (i.e, company called).

Table 1: A collection of service ticket summaries

Objects	Condition , attributes				
Ticket ID	**Agent ID**	**Product ID**	**Complaint ID**	**Service type**	**Decision Attribute**
F1	T01	30	1	Outbound	Failure
F2	T01	30	1	Outbound	Success
F3	T01	31	1	Outbound	Failure
F4	T01	31	2	Outbound	Failure
F5	T01	31	3	Outbound	Failure
F6	T01	31	1	Inbound	Success
F7	T01	31	2	Inbound	Success
F8	T01	31	3	Inbound	Success
F9	P02	31	1	Outbound	Success
F10	P02	31	2	Outbound	Success
F11	P02	31	3	Outbound	Success

In Table 1 each ticket summary is one object (i.e., one record); the condition attributes that are defined earlier are Agent ID, Product ID, Complaint type, and Service type; and the decision attribute which has only two possible values: Failure and Success. Elementary sets can be obtained by simply listing the number of failures and successes in the D-space (decision):

Set Failure: {F1, F3, F4, F5} (total 4), Set Success: {F2, F6, F7, F8, F9, F10, F11} (total 7)

6. CALCULATION OF UPPER AND LOWER APPROXIMATIONS FROM THE ELEMENTARY SETS

In our decision Set Failure, F1 and F2 have all the same condition attributes values but with different decision (F1 has failure but in F2 there is success) so, this condition set will be regarded as uncertain and it will be ignored from the calculation of the lower approximation; but we will include it in the calculation of upper approximation. Thus: lower approximation = {F3, F4, F5} (Failure set), upper approximation = {F1, F2, F3, F4, F5} (Failure set). The boundary region = {F1, F2} (this can be classified either as Failure or not-Failure (Success)). Using the same to concept of Success, we find the objects to have the lower approximation = {F6, F7, F8, F9, F10, F11}, the upper approximation {F1,F2,F6,F7,F8,F9,F10,F11}, and the boundary region we get is = {F1,F2}. The quality of lower approximation is (6 + 3)/(7 + 4), or 0.8182; the accuracy for success is 0.75 (6/8); the accuracy for Failure is 0.60 (3/5); and the accuracy of the whole classification is (6 + 3)/(8 + 5), or 0.6923.

The accuracy of our classification using all condition attributes sets is given by:

	failure	Success
Number of records	4	4
Number of lower approximation	3	6
Number of upper approximation	5	8
Accuracy	0.6	0.75

7. ISSUES WITH THE DECISION TABLE

Same or indisernable objects may be represented many times and some of the attributes may be superfluous (redundant). That is, their removal cannot affect the classification.

7.1 Finding reducts and core

Attributes that have indiscernibility relation are required to frame up a set approximation. There are several such subsets of attributes present which are minimal and are been categorised as 'reducts'. A reduct, actually, is the minimal subset of all attributes that enables the same classification of elements of the universe as the whole set of attributes and attributes that do not belong to a reduct are superfluous with respect to classification of elements of the universe. Now the question arises, how can we remove some data from a data table containing some superfluous data. For example, it is easily seen that if we drop in Table 1 either the Complaint Type and Product ID, we get the data set which is equal to the original one with regard to approximations and dependencies. Reducts which will be found - {Agent ID, Product ID, Service type} and {Agent ID, Complaint ID, Service Type} . Thus we get the same accuracy of approximation and degree of dependencies as it is given in the table 1. If B be a subset of A then ‘a’ belong to B; we can depict that ‘a’ is dispensable in B if I(B) = I(B – {a}), otherwise a is indispensable in B. Set B is found out to be independent if all its attributes are indispensable. So, Set B' of B is a reduct of B if B' is independent and I(B') = I(B). If all its attributes of set B are indispensable then B is independent.

Now we have to define the core of attributes. Let B be a subset of A. The core of B is the set of all indispensable attributes of B. The following is an important property, connecting the core and the reduct $CORE(B) = \bigcap RED(B)$, where RED(B) is the set off all reducts of B, the core is the intersection of all reducts and will include in every reduct. Therefore, the core is an important subset of attributes. We know that, each row of a decision table represents a decision rule. In table 1 decision rules 1 and 2, both have all the same conditions but different decision type. Such rules are inconsistent otherwise rules are referred as consistent. In Table 1, there are two relative reducts for decision type{Agent ID, Product ID, Service Type}and {Agent ID, Complaint ID, Service Type} from the set of conditional attributes { Agent ID ,Complaint ID, Product ID, Service Type}.As we know core is intersection of both {Agent ID, Product ID, Service Type}and {Agent ID, Complaint ID, Service Type} , so after taking the intersection of attributes we get the core attributes that is { agent ID , service type}, and we can eliminate the rest of attributes{complaint ID ,product ID}.

8. REDUCTION OF ATTRIBUTE

So we have find out one attribute which is superfluous here, that is Complain ID. Now we remove it from the table 1.

Table 2: Reduced table

Ticket ID	Agent ID	Product ID	Service type	Decision Type
F1	T01	30	Outbound	Failure
F2	T01	30	Outbound	Success
F3	T01	31	Outbound	Failure
F4	T01	31	Outbound	Failure
F5	T01	31	Outbound	Failure
F6	T01	31	Inbound	Success
F7	T01	31	Inbound	Success
F8	T01	31	Inbound	Success
F9	P02	31	Outbound	Success
F10	P02	31	Outbound	Success
F11	P02	31	Outbound	Success

In the same way if we remove Product ID, we will find that we can still get the same result by not affecting our decision.

Table 3: Reduced table

Agent ID	Service type	Decision Type
T01	Outbound	Failure
T01	Outbound	Success
T01	Outbound	Failure
T01	Outbound	Failure
T01	Outbound	Failure
T01	Inbound	Success
T01	Inbound	Success
T01	Inbound	Success
P02	Outbound	Success
P02	Outbound	Success
P02	Outbound	Success

Finally, after further reduction (after removal of attributes and attributes values), we get the above table. It is clear from the table that we get the same result after the reduction also and the data has got reduced and the following table is the ultimate version of the Table 1.
After further investigation, we now have a procedure for eliminating values of attributes from table 3 that do not influence on decision.

Tavle 4: Reduced table of table 3

Agent-ID	Service Type	Decision-Type
T01	Outbound	Failure
T01	Outbound	Success
•	Inbound	Success
•	Inbound	Success
•	Inbound	Success
P02	•	Success
P02	•	Success
P02	•	Success

Here, the attributes needed for the classification has been reduced. Along with the removal of Complaint Type and Product Id all reducts are reduced.
Thus superfluous data is removed from our table. Therefore, our optimized result in table 4 can designed as a decision algorithm.

Decision algorithm:

If(Agent ID = T01)and (Service Type = Outbound)then(decision= Success),
If(Agent ID = T01)and (Service Type = Outbound)then(decision= failure),
If(Service type = Inbound)Then(decision = Success),
If(Agent ID = P02)Then(decision = Success).

CONCLUSIONS

In the above article we have showcased the application of the data mining procedure based upon rough set theory through the usage of core and reducts after collecting some relevent data from call centre. It's a logical approach in data sets which is been used to draw a conclusion from factual data. After finding core and reduct we have succeded in data reduction by reducing the number of attributes, and attributing values without disturbing the accuracy. Undoubtedly, it's been a successful attempt in producing an effective algorithm based upon the above set and such application of rough set will surely increase the efficiency of all such organisations which deal with the task of data restoration.

REFERENCES

[C+11] **Rong-Rong Chen, Yen-I. Chiang, P. Pete Chong, Yung-Hsiu Lin, Her-Kun Chang** - *Rough set analysis on call center metrics*, Applied Soft Computing 11, pp.3804–3811, 2011

[Paw01] **Zdzisław Pawlak** - *Rough sets and intelligent data analysis*, 2001

[SD01] **J. Saquer, J. S. Deogun** - *Concept approximations based on rough sets and similarity measures*, International Journal of Applied Mathematics and Computer Science 11 (3), 655–674, 2001