Rough Set Applications for the Classification of Managerial Policies Effect on Industries

Ratnakar Das
Research Scholar, BPUT

Deepti Mishra
CET, Bhubaneswar

Sujogya Mishra
Research Scholar, Utkal University

ABSTRACT
Rough Set is a very handy soft computing tools for data analysis. This paper analyze the policies of the managements towords their employees using RST

Keywords
RST-Rough Set Theory

1. INTRODUCTION
The concept of RST was proposed by Zdzislaw Pawlak in the early 1980[2] is a handy tool to deal with voluminous data analysis. Rough set theory can be applied on various fields like data analysis and data mining. This approach (RST) is fundamental for artificial intelligence (AI) and cognitive sciences, particularly in the fields of ML, KA, DA, and KD from database, DSS, IR, and PR[2,5,9]. Suppose the information system in us is E=(U , A), X $\subset$ U and P $\subset$ A,where U and A are finite, nonempty sets called the universe, and the set of attributes, respectively. Set A contains two disjoint sets of attributes called condition and decision attributes and the system is represented by

S = (U, C, D) where C is called condition attribute and D is called decision attribute. With every attribute aϵ A we associate a set Va, of its values, called the domain of a.

The RST concept deals with Lower and Upper approximation these two concepts are the backbone of RST in Rule derivation . Upper approximation and Lower approximation denoted as $\overline{p(x)}$, $\underline{p(x)}$ respectively.Where both approximations are defined as follows

$$\underline{P(x)} = \{ \forall x \in U; P(x) : P(x) \subset X \}$$

$$\overline{P(X)} = \{ \forall x \in U; P(x) : P(x) \cap X \neq \varphi \}.$$

Initially considered 1000 raw samples of industries which are sick due to their productivity. we consider managerial policies as major in this paper. Applying clustering techniques the 1000 samles size reduced to 6 different clusters then using roush set concept we are able to find the important managerial attributes to increase the productivities .

(ML-Machine Learning,KA-Knowledge Acquisition,DA-Decesion Analysis,KD- Knowledge Discovery,DSS-Decesion Support System,IR-Inductive Reasoning,PR-Pattern Recognition ,RST-Rough Set Theory)

1.1 Basic Ideas
The basic idea was developed looking at the present situation of industries affected by managerial policies. Our basic concept to find the attributes responsible for industries(heavy). Initially considered 6 clusters {R_1,R_2,R_3,R_4,R_5,R_6} as its records and the conditional attributes are working enivironments renamed as t_1,Relation between manager and their subordinates renamed as t_2,Quality equipmens as t_3, Sufficient resting time for ground label workers as t_4, Sudden increase of production more than the per day capacity as t_5 and low quality machine t_6 . It's values which are significant or insignificant renamed as p_1 and p_2 respectivly. The decision attribute is renamed as r and their values are success and failure renamed as k_2 and k_1 respectivly.

The paper is organized in the following manner, section -1: about introduction, section -2 : about basic ideas to find core and reduct by using quick reduct algorithm, section-3: finding reduct using strength and section-4: about Experiment and Conclusion.

1.2 Application of Cluster in Data Analysis
Cluster analysis or clustering is the technique of classifying a set of entire object space in such a way that in the same group elements are more similarity (with respect to each other). It is basically used in data mining and in statistical data analysis. It can be attained by using different algorithms that differ significantly with respect to run time and space complexities.

1.3 Types of Clustering Algorithms
1. Partitioning-based clustering algorithms are that determine all the clusters at once in most cases.
   - K-means clustering
   - K-medoids clustering
   - EM (expectation maximization) clustering
2. Hierarchical clustering: these algorithms find successive clusters using previously established ones.
   - Divisive clustering is a top down approach.
   - Agglomerative clustering is a bottom up approach.
3. Density-Based Methods: these clustering algorithms are used to help discover arbitrary-shaped clusters. A cluster is defined as a region in which the density of data objects exceeds some threshold.

Algorithmic steps for k-means clustering

Let $Y = \{y_1, y_2, y_3, \ldots\ldots\ldots, y_n\}$ be the set of data points and $W = \{w_1, w_2, \ldots\ldots\ldots, w_c\}$ be the set of centers.

1) Randomly select 'c' cluster centers.

2) Calculate the distance between each data point and cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..

4) Recalculate the new cluster center using:

$w_i = \sum_{i=1}^{c_i} y_i$ where, 'c_i' represents the number of data points in i^{th} cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.

6) If no data point is reassigned then stop, otherwise repeat from step 3.

2. DATA REDUCTION USING DECISION TABLE

To find the attributes responsible in establishing the Software industries , the information table is formed by using the collected data about the Software industries from different sources . The Information table-1 presented below. Using quick reduct algorithm on the collected data it found a set of reduct .

InformationTable-1

E	t_1	t_2	t_3	t_4	t_5	t_6	r
R_1	p_1	p_2	p_2	p_1	p_1	p_2	k_2
R_2	p_1	p_2	p_1	p_1	p_2	p_2	k_2
R_3	p_1	p_1	p_2	p_1	p_1	p_2	k_1
R_4	p_2	p_1	p_1	p_2	p_2	p_1	k_1
R_5	p_2	p_1	p_1	p_2	p_1	p_2	k_1
R_6	p_2	p_1	p_1	p_2	p_2	p_2	k_1

From the above information tables we have the following sets of Reduct[9] that are as follows

1 (t_1, t_2, t_3, t_4)
2. (t_2, t_3, t_4, t_5)
3. (t_1, t_2, t_4, t_6)
4. (t_1, t_2, t_4, t_5)
5. (t_2, t_3, t_4, t_6)

From the above information table here are the 5 sets of reduct to calculate core by using these reducts. The result as follows: Core = $\bigcap$ Reduct i.e . (t_2, t_4) as core set to verify this. The concept of strength to find the reduct and core is given in the following sub sections.

3. FINDING REDUCT USING STRENGTH OF ROUGH SET

Same data set is being implemented by using strength (Rough set theory) and again 6 samples are considered for the application of streangth of RST which is obtained by statistical correlation techniques taking 1000 samples, 6 conditional attributes and two decision attributes same as the information table-1 stated above and the Proposed Algorithm is as follows

1. begin

2. Initialize Reduct set as $k = \varphi$

3. do N (attribute sets) N and K $\neq \varphi$ for all attributes

(conditional attribute values with respect to decision attribute values)

4. Continue to find Equivalence classes by using

Strength=(conditionalattribute)value/(Decision attribute)value = D-values/C-values,

where D-values: decision attribute values and C-values: conditional attribute values i.e with respect to cardinality of both conditional attribute values and decision attribute values

5. If ratio count of conditional attribute values to decision

attribute values, falls in a group , Reduct++

else goto step 4

end{if}

end{for}

while no further classification is possible

Where N & K $\in$ E(Records).

Information Table-2

E	t_1	t_2	t_3	t_4	t_5	t_6	r
R_1	p_1	p_2	p_2	p_1	p_1	p_2	k_2
R_2	p_1	p_2	p_1	p_1	p_2	p_2	k_2
R_3	p_1	p_1	p_2	p_1	p_1	p_2	k_1
R_4	p_2	p_1	p_1	p_2	p_2	p_1	k_1
R_5	p_2	p_1	p_1	p_2	p_1	p_2	k_1
R_6	p_2	p_1	p_1	p_2	p_2	p_2	k_1

Meaning of $(t_1, t_2, t_3, t_4, t_5, t_6)$, (p_1, p_2) and (k_1, k_2) are described in the above section.

Here in this case our target is to find the reduct using the strength $R_{Sucess} = \{ R_1, R_2 \}$

$R_{Failure} = \{ R_3 , R_4, R_5, R_6\}$ now finding $R_{Sucess}(t_1)$ p_1=33%, $R_{Failure}(t_1)p_2$=Nil , similarly finding $R_{Sucess}(t_2)lp_1$=100%, $R_{Failure}(t_2)p_2$=100% , $R_{Sucess}(t_3)p_1$=75%, $R_{Failure}(t_3)p_2$=50%, $R_{Sucess}(t_4)p_1$=33%, $R_{Failure}(t_4)p_2$=Nil, %, $R_{Sucess}(t_5)p_1$=66%, $R_{Failure}(t_5)p_2$=33%, $R_{Sucess}(t_6)p_1$=100% $R_{Failure}(t_6)$ p_2=25%

From the above analysis it is clear that attribute t_1,t_6 produces extreme result so we drop both attributes from the Information Table-2 leads to next Table-3

Reduct Table-3

E	t_2	t_3	t_4	t_5	r
R_1	p_2	p_2	p_1	p_1	t_2
R_2	p_2	p_1	p_1	p_2	t_2
R_3	p_1	p_2	p_1	p_1	t_1
R_4	p_1	p_1	p_2	p_2	t_1
R_5	p_1	p_1	p_2	p_1	t_1
R_6	p_1	p_1	p_2	p_2	t_1

Upon analyzing Table-3 we have the following result i.e. { R_4, R_6} produces same result so merge the two fields in to one field so new table found as follows.

Reduct Table-4

E	t_2	t_3	t_4	t_5	r
R_1	p_2	p_2	p_1	p_1	t_2
R_2	p_2	p_1	p_1	p_2	t_2
R_3	p_1	p_2	p_1	p_1	t_1
R_4	p_1	p_1	p_2	p_2	t_1
R_5	p_1	p_1	p_2	p_1	t_1

Information table-4 cannot further classified. Rule generated from table-4 as follows

- t_2(insignificant),t_3(insignificant),t_4(significant),t_5(significant) →Success
- t_2(insignificant),t_3(significant),t_4(significant),t_5(insignificant) → Success
- t_2(significant),t_3(insignificant),t_4(significant),t_5(significant) →Failure
- t_2(significant),t_3(significant),t_4(insignificant),t_5(insignificant) →Failure
- t_2(significant),t_3(significant),t_4(insignificant),t_5(significant) →Failure

4. EXPERMENTAL SECTION

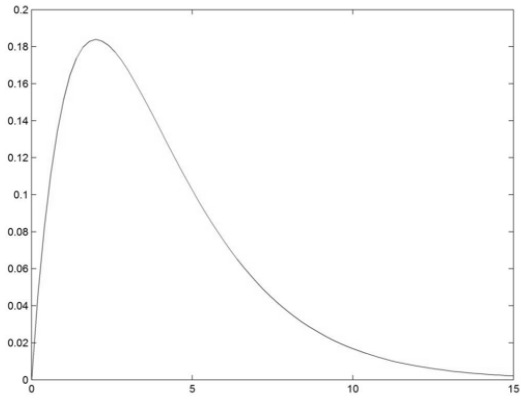

Figure-1

Industries

	A	B	C
	643	469	484
	655	427	456
	702	525	402
$\overline{X}$	666.67	473.67	447.33
S	31.18	49.17	41.68

Null Hypothesis :- $\mu_0 = \mu_1 = \mu_2$ there is no impact of attributes in increasing the productivity

Alternate Hypothesis :- $\mu_0 \neq \mu_1 \neq \mu_2$ there is a increase in productivity with satisfying attributes

We get a F = $\frac{MSTR}{MSE} = \frac{43024.78}{1709} = 25.17$ that is much larger than the critical values 5.14 so we reject the null hypothesis and accept the alternate one i.e $\mu_0 \neq \mu_1 \neq \mu_2$ there is a increase in productivity with satisfying attributes the above table is an anova table .

5. CONCLUSION

This paper we generalize the RST concept by two way analysis one by using quick reduct algorithm anthor one using streangth of rough set futher this concept of comparative studies can be extended to many different types of domain.

6. REFERENCES

[1] S.K. Pal, A. Skowron (Eds.), Rough Fuzzy Hybridization, Springer, Berlin, 1999

[2] Z. Pawlak, Rough Sets, International Journal of Computer and Information Sciences, 11 (1982) 341–356.

[3] Z. Pawlak, Rough Sets: Theoretical Aspects of Reasoning about Data, System Theory, Knowledge Engineering and Problem Solving, 9, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1991

[4] Z. Pawlak, Decision Rules, Bayes_ rule and Rough Sets, in: N. Zhong, A. Skowron, S. Ohsuga (Eds.), New Direction in Rough Sets, Data Mining, and Granular-Soft Computing, Springer, Berlin, 1999, pp. 1–9

[5] L. Polkowski, A. Skowron (Eds.), "Rough Sets and Current Trends in Computing", Lecture Notes in Artificial Intelligence, 1424, Springer, Berlin, 1998

[6] L. Polkowski, A. Skowron (Eds.), "Rough Sets in Knowledge Discovery", 1–2, Physica Verlag, A Springer Company, Berlin, 1998.

[7] L. Polkowski, S. Tsumoto, T.Y. Lin (Eds.), "Rough Set Methods and Applications–New Developments in Knowledge Discovery in Information Systems", Springer, Berlin, 2000, to Appear.

[8] N. Zhong, A. Skowron, S. Ohsuga (Eds.), New Direction in Rough Sets Data Mining and Granular-Soft Computing, Springer, Berlin, 1999.

[9] Renu Vashist M.L.Garg "Rule Generation based on Reduct and Core:A Rough Set Approach",vol-29 no-9 IJCA 2011