Rough Set Applications for the Classification of Software Industries using Rule based Approach

Ratnakar Das
Research Scholar, BPUT

Deepti Mishra
CET, Bhubaneswar

Sujogya Mishra
Research Scholar, Utkal University

ABSTRACT
Rough Set theory is a very handy tool for imprecise and vague pattern of data . This paper shows how the concept of RST being used in deriving information from hidden pattern of data . From large data base software industries are the object of interest for applying Rough Set concept on the collected data. The set of rules which have been derived will be helpful in the development of software industries. This paper has used two types of techniques in finding the reduct, first one uses cluster in finding different dissimilar groups the other one is the application of quick reduct algorithm in deriving the rules verifying them by using strength.

Keywords
RST-Rough Set Theory,

1. INTRODUCTION
The concept of RST developed by Zdzislaw Pawlak in the early 1980[2] is a new mathematical tool to deal with vagueness and uncertainty. Rough set theory is applied on various field like data analysis and data mining. This approach (RST) is fundamental for artificial intelligence (AI) and cognitive sciences, particularly in the fields of machine learning, knowledge acquisition, decision analysis, and knowledge discovery from database, expert systems, decision support systems, inductive reasoning, and pattern recognition[2,5,9]. Rough set theory can be considered as a particular case of classical set theory. RST can be used in extracting strong conclusion from incomplete information. The basic concept of RST is the notion of approximation space with every object of universe to associate with some information i.e. Data and Knowledge.

Suppose we are given an information system 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

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

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

Lower approximation consists of all the members which surely belongs to the set and Upper approximation consists of all the members which possibly belong to the set as defined above.

Information table is the most important aspect of RST as rows of the information table are called records and columns are called conditional and decision attributes .

Using the above concept a set of rule is derieved for the software industries , initially 1000 samples were collected from sick software industries, then applied clustering techniques the sample size reduced to 6 different clusters , applying rough set concept on those it found the important attributes responsible to establish software industries . This paper has applied two different concepts of finding reduct . 1st approach is the use of quick reduct algorithm and the 2nd approach is to find reduct by using strength where data set is grouped by using correlation techniques.

1.1 Basic ideas
The basic idea was developed looking at the present situation of Software industries .The major intention is to find which attributes are responsible in establishing the Software industries. Initially considered 6 clusters $\{E_1,E_2,E_3,E_4,E_5,E_6\}$ as its records and the conditional attributes are Location for Industries renamed as a_1,quality technician which includes Software Engineers renamed as a_2, Good and organized work culture renamed as a_3, High quality equipments and strong research and analysis wings as a_4, strong group of Industrial administrators who can properly present the company around the globe are renamed as a_5 and presence of good and qualitative post-operational facilities renamed as a_6 . It's values which are significant or insignificant renamed as b_1 and b_2 respectivly. The decision attribute is renamed as d 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.

1.4 Algorithmic steps for k-means clustering

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

Information Table-1

E	a_1	a_2	a_3	a_4	a_5	a_6	d
E_1	b_1	b_2	b_2	b_1	b_1	b_2	k_2
E_2	b_1	b_2	b_1	b_1	b_2	b_2	k_2
E_3	b_1	b_1	b_2	b_1	b_1	b_2	k_1
E_4	b_2	b_1	b_1	b_2	b_2	b_1	k_1
E_5	b_2	b_1	b_1	b_2	b_1	b_2	k_1
E_6	b_2	b_1	b_1	b_2	b_2	b_2	k_1

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

1 (a_1,a_2,a_3,a_4)
2. (a_2,a_3,a_4,a_5)
3. (a_1,a_2,a_4,a_6)
4. (a_1,a_2,a_4,a_5)
5. (,a_2,a_3,a_4,a_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 . it has (a_2,a_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	a_1	a_2	a_3	a_4	a_5	a_6	d
E_1	b_1	b_2	b_2	b_1	b_1	b_2	k_2
E_2	b_1	b_2	b_1	b_1	b_2	b_2	k_2
E_3	b_1	b_1	b_2	b_1	b_1	b_2	k_1
E_4	b_2	b_1	b_1	b_2	b_2	b_1	k_1
E_5	b_2	b_1	b_1	b_2	b_1	b_2	k_1
E_6	b_2	b_1	b_1	b_2	b_2	b_2	k_1

Meaning of (a_1, a_2 , a_3 , a_4 , a_5 , a_6) , (b_1, b_2) and (d_1, d_2) are described in the above section.

HHere in this case our target is to find the reduct using the strength E_{Sucess} ={ E_1, E_2 }

$E_{Failure}$ ={ E_3 , E_4, E_5, E_6} now finding $E_{Sucess}(a_1)$ b_1=33%, $E_{Failure}(a_1)$ b_2=Nil , similarly finding $E_{Sucess}(a_2)$ b_1=100%, $E_{Failure}(a_2)$ b_2=100% , $E_{Sucess}(a_3)$ b_1=75%, $E_{Failure}(a_3)$ b_2=50%, $E_{Sucess}(a_4)$ b_1=33%, $E_{Failure}$ (a_4) b_2=Nil, %, $E_{Sucess}(a_5)$ b_1=66%, $E_{Failure}(a_5)$ b_2=33%, $E_{Sucess}(a_6)$ b_1=100% $E_{Failure}(a_6)$ b_2=25%

From the above analysis it is clear that attribute a_1,a_6 produces extreme result so we drop both attributes from the Information Table-2 leads to next Table-3

Reduct Table-3

E	a_2	a_3	a_4	a_5	d
E_1	b_2	b_2	b_1	b_1	k_2
E_2	b_2	b_1	b_1	b_2	k_2
E_3	b_1	b_2	b_1	b_1	k_1
E_4	b_1	b_1	b_2	b_2	k_1
E_5	b_1	b_1	b_2	b_1	k_1
E_6	b_1	b_1	b_2	b_2	k_1

Upon analyzing Table-3 we have the following result i.e. { E_4, E_6} produces same result so merge the two fields in to one field so new table found as follows.

Reduct Table-4

E	a_2	a_3	a_4	a_5	d
E_1	b_2	b_2	b_1	b_1	k_2
E_2	b_2	b_1	b_1	b_2	k_2
E_3	b_1	b_2	b_1	b_1	k_1
E_4	b_1	b_1	b_2	b_2	k_1
E_5	b_1	b_1	b_2	b_1	k_1

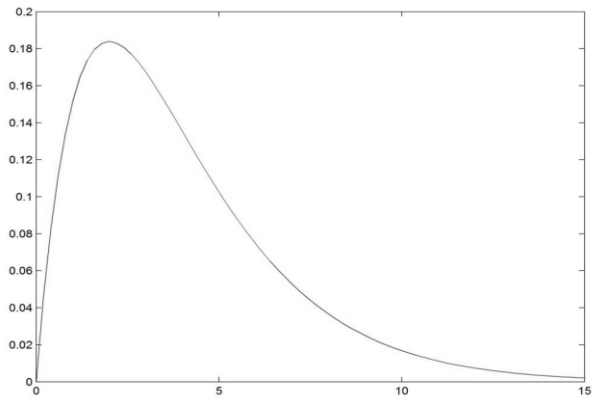

Figure-1

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

- a_2(insignificant),a_3(insignificant),a_4(significant),a_5(significant) →Success
- a_2(insignificant),a_3(significant),a_4(significant),a_5(insignificant) → Success
- a_2(significant),a_3(insignificant),a_4(significant),a_5(significant) →Failure
- a_2(significant),a_3(significant),a_4(insignificant),a_5(insignificant) →Failure
- a_2(significant),a_3(significant),a_4(insignificant),a_5(significant) →Failure

4. EXPERMENTAL SECTION

χ^2 distribution is used for this purpose. The samples are collected from different sources with Expectation 15%,10%,15%,20%,30%,15% and the Observed samples are 25,14,34 45,62,20. It has total 200 samples, so expected number of samples per each day is as follows 30,20,30,40,60,30. Then chi square distribution has been applied to verify the result assuming that H_0 is the hypothesis that is correct H_1 as alternate hypothesis that is not correct , Then we expect sample in six cases as chi squared estimation formula is $\sum(O_i-E_i)^2/ E_i$ where i=0,1,2,3,4,5 so the calculated as follows

χ^2 =(25-30)^2/20+(14-20)^2/20+(34-30)^2/30+(45-40)^2/40+(62-60)^2/60+(20-30)^2/30=7.60 this is much below tabular values i.e. 11.04.Figure using mat-lab provided below

5. CONCLUSION

This paper has summed up that both Quality technicians which include Software Engineers and High quality equipments, strong research and analysis wings lead to success in software business by Quick Reduct method and set of rules are generated by using strength of RST. Both the methods provide a predictive analysis of attribute reduction.

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