Rough set (RS) approach for optimal rule generation in medical data

Prof. Dr. P K Srimani
Former Chairman, Dept. of CS & Maths,
Bangalore University, Director, R&D, B.U.,
Bangalore, India

Manjula Sanjay Koti
Assistant Professor, Dept. of MCA, Dayananda Sagar
College of Engineering, Bangalore,
Research Scholar, Bharathiar University, Coimbatore, India.

Abstract— **Medical data mining has great potential for exploring the hidden patterns in the data sets of the medical domain and these patterns can be utilized for clinical diagnosis. However, the available raw medical data are widely distributed, heterogeneous in nature, and voluminous. Medical databases have accumulated large quantities of information about patients and their medical conditions. Relationships and patterns within this data could provide new medical knowledge. Analysis of medical data is often concerned with the treatment of incomplete knowledge, with management of inconsistent pieces of information and with manipulation of various levels of representation of data. In the present study, the theory of RS is applied to find dependence relationship among data, evaluate the importance of attributes, discover the patterns of data, learn common decision-making rules, reduce all redundant objects and attributes and seek the minimum subset of attributes so as to attain satisfactory classification. It is concluded that the decision rules(with and without reducts) generated by the rough set induction algorithms(Exhaustive, Covering and LEM2) not only provide new medical insight but also are useful for medical experts to analyze the problem effectively.**

Keywords- Reducts, Rule generation, Rough set, Medical mining, Induction Algorithms.

I. INTRODUCTION

Data mining is an essential process of applying intelligent methods in order to extract data patterns, pattern evaluation to identify the truly interesting patterns based on some interesting measures and knowledge presentation which uses visualization and knowledge representation for presenting the mined knowledge to the user[1]. The process of finding useful patterns or meaning in raw data has been called knowledge discovery in databases [2]. Medical data mining has great potential for exploring the hidden patterns in the data sets of the medical domain and these patterns can be utilized for clinical diagnosis. However, the available raw medical data are widely distributed, heterogeneous in nature, and voluminous. These data need to be collected in an organized form. This collected data can be then be integrated and made available to a Hospital Information System (HIS).Actually, medical databases have accumulated large quantities of information about patients and their medical conditions. Relationships and patterns within this data could provide new medical knowledge. Analysis of medical data is often concerned with the treatment of incomplete knowledge, with management of inconsistent pieces of information and with manipulation of various levels of representation of data.

The success of machine learning associated with medical data sets is strongly affected by many factors and one such factor is the quality of the data which depends on irrelevant, redundant and noisy data. Thus, when the data quality is not excellent, the prediction of knowledge discovery during the training process becomes an arduous task. The existing intelligent techniques [4,5] of medical data analysis are concerned with (i) Treatment of incomplete knowledge (ii) Management of inconsistent pieces of information and (iii) Manipulation of various levels of representation of data. This difficulty is minimized by feature selection which identifies and removes the irrelevant and redundant features in the data to a great extent.

Further, Intelligent methods such as neural networks, fuzzy sets, decision trees and expert systems are applied to the medical fields[6,7] but cannot derive conclusions from incomplete knowledge or can manage inconsistent information. A proper selection of a subset of attributes/features is absolutely essential to represent the patterns that are to be classified in problems like practical pattern classification and knowledge discovery problems.

In recent years Classical rough set theory developed by Professor Z. Pawlak in 1982 has made a great success in the field of knowledge acquisition [3]. A fundamental principle of a rough set based learning system is to discover redundancies and dependencies between the given features of a problem to be classified. It approximates a given concept below and from above, using lower and upper approximations. Consequently, a rough set learning algorithm can be used to obtain a set of rules in IF-THEN form, from a decision table. The theory of RS can be used to find dependence relationship among data, evaluate the importance of attributes, discover the patterns of data, learn common decision-making rules, reduce all redundant objects and attributes and seek the minimum subset of attributes so as to attain satisfactory classification. We have used Pima data set for our study, which has been widely used in machine learning experiments and is currently available through the UCI repository of standard data sets. The present investigation on Rough sets is organized as follows: Related

work, Methodology, Experiments and Results. Finally the conclusions are presented.

A. Applications

Rough sets have been proposed for a very wide variety of applications. In particular, the rough set approach seems to be important for Artificial Intelligence and cognitive sciences, especially in machine learning, knowledge discovery, data mining, expert systems, approximate reasoning and pattern recognition. Rough set rule induction algorithms generate decision rules [8,9], which not only provide new medical insight but also are useful for medical experts to analyze the problem effectively. These decision rules are more useful for medical experts to analyze and gain understanding into the problem at hand.

II. LITERATURE SURVEY

Tsumoto [10] proposed a rough set algorithm to generate diagnostic rules based on the hierarchical structure of differential medical diagnosis. The induced rules can correctly represent experts' decision processes. Komorowski and Ohrn [11] use a rough set approach for identifying a patient group in need of a scintigraphic scan for subsequent modeling. Bazan [12] compares rough set-based methods, in particular dynamic reducts, with statistical methods, neural networks, decision trees and decision rules. He analyzes medical data, i.e. lymphography, breast cancer and primary tumors, and finds that error rates for rough sets are fully comparable as well as often significantly lower than that for other techniques. In Ref. [13,14], a rough set classification algorithm exhibits higher classification accuracy than decision tree algorithms. The generated rules are more understandable than those produced by decision tree methods. Some of the other works include [15,16].

III. DATA SET DESCRIPTION

We have used Pima data set for our study, which has been widely used in machine learning experiments and is currently available through the UCI repository of standard data sets. To study the positive as well as the negative aspects of the diabetes disease, Pima data set can be utilized, which contains 768 data samples. Each sample contains 8 attributes which are considered as high risk factors for the occurrence of diabetes, like Plasma glucose concentration, Diastolic blood pressure (mm Hg), Triceps skin fold thickness (mm), 2-hour serum insulin (mu U/ms), Body mass index (weight in kg/(height in m)) Diabetes pedigrees function and Age (years). All the 768 examples were randomly separated into a training set of 576 cases (378, non-diabetic and 198, diabetic) and a test set of 192 cases (122 non-diabetic and 70 diabetic cases).

IV. METHODOLOGY

The present study illustrates how set theory could be used for the analysis of medical data especially for generating classification rules from a set of observed samples of the pima data set. In rough set theory, knowledge is represented in information systems. An information system is a data set represented in a tabular form called decision table in which each row represents an object (for eg. a case or an event) and each column represents an attribute. In order to determine all the reducts of the data that contains the minimal subset of attributes that are associated with a class label for classification. The Rough Set reduction technique is employed. In a knowledge system reducts are often used at the data preprocessing stage during the attribute selection process. It is important to note that reduct is not unique and in a decision table multiple reducts may exists. The core of a decision table which consists of essential information is certainly contained in every reduct. In other words, a reduct generated from the original data set should contain the core attributes. During the attribute selection process reduct and core are the commonly used since the main purpose of rough set theory is to select the most relevant attributes with regard to the classification task and to remove the irrelevant attributes. The set of attributes which is common to all reducts is called the core: which is possessed by every legitimate reduct, and hence consists of essential attributes which cannot be removed from the information system without causing collapse of the equivalence-class structure. In other words, a core is absolutely necessary for the representation of the categoric structure.

Rough set is used to derive the classification rules in the medical data. The key features of Rough sets are:

(i) It does not need any preliminary or additional information about data – like probability in statistics, grade of membership in the fuzzy set theory

(ii) It provides efficient methods, algorithms and tools for finding hidden patterns in data.

(iii) It allows to reduce original data, i.e. to find minimal sets of data with the same knowledge as in the original data

(iv) It allows to evaluate the significance of data

(v) It allows to generate in automatic way the sets of decision rules from data

(vi)It is easy to understand.

(vii) It offers straightforward interpretation of obtained results

(viii)It is suited for concurrent (parallel/distributed) processing

(ix) It has easy internet access to the rich literature about the rough set theory, its extensions as well as interesting applications.

A. Rule Induction

It is emphasized that the number of all minimal consistent decision rules for a given decision table can be exponential with respect to the size of decision table. Three heuristics have been implemented in RSES:

B. Exhaustive Algorithm

This algorithm realizes the computation of object oriented reducts (or local reducts). It has been shown that some minimal consistent decision rules for a given decision table S can be obtained from objects by reduction of redundant descriptors. The method is based on Boolean reasoning approach.

```
exhaustive(int sol, int depth)
{
    if (issolution(sol))
        printsolution(sol)
     else
     {
        solgenerated = generatesolution()
        exhaustive(solgenerated,depth+1)
}
```

C. Covering Algorithm

This algorithm searches for minimal (or very close to minimal) set of rules which cover the whole set of objects.

```
Inputs: labeled training dataset D

Outputs: ruleset R that covers all instances in D
Procedure:
      Initialize R as the empty set
      for each class C {
            while D is nonempty {
       Construct one rule r that correctly classifies some
       instances in D that  belong to class C and does not
       incorrectly   classify any non-C instances
        Add rule r to ruleset R
        Remove from D all instances correctly classified by r
              }
              }
      return R
```

D. LEM2 Algorithm

This is a realization of LEM2 algorithm, which is another kind of covering algorithm **(see [4]).**

```
Input: B set of objects
Output: R set of rules
begin
G = B;
R = φ;
While G ≠φ do
begin
C ≠φ
C(G) = {c: [c]∩G≠φ};
While(C ≠φ) or (!([C] ⊆ B)) do
begin
select a pair c ∈C(G) such that |[c] ∩ G| is maximum;
if ties, select a pair c ∈C(G) with the smallest cardinality
|[c]|;
if further ties occur, select the first pair from the list;
C = C ∪{c};G = [c]∩G;
C(G) = {c:[c]∩G≠φ};
C(G) = C(G) – C;
end;
for each elementary condition c ∈ C do
if |C – c| ⊆ Bthen C = C – {c};
create rule r basing the conjuction C and add it to R;
G == B –U   |R|
r ∈ R
end;
for each r ∈ R do
if  U  |S|=B then R = R –r
s ∈ R- r
end.
```

The Rough set philosophy was founded on assumption that every object of the universe set associated with some information (knowledge, data) [4] . All objects with similar information are indiscernible and form blocks, which can be considered as elementary granules. These granules are called concepts and can be considered as elementary granules. These granules are called concepts and can be considered as elementary building blocks of our knowledge. Any union of elementary sets is called a crisp, and any other sets are referred to as rough(vague) .Consequently each rough set has boundary line, which is the objects that cannot be with certainly classified as members of the set or of it's complement. Fig.2 shows the mining methodology of the patient data using rough set theory.

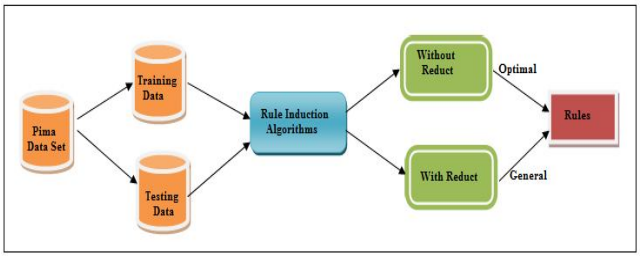

Fig 2. Mining the patient data using rough set

In order to reduce the redundancy and irrelevancy/inconsistency present in the attributes, the concept of decision rule generation through reducts has emerged in Rough set theory. Actually, a reduct is a subset of conditional attributes representing the entire data table. The determination of suitable reducts is a challenging problem and researchers are working hard in this direction to find suitable algorithms. The decision rules will be generated from reducts and could be used for the classification of objects.

The experiment was conducted on Pima Indian Diabetes dataset which contains 768 samples with two-class problem. The problem posed here is to diagnose whether a patient would test positive or negative for diabetes. The diagnosis can be carried out based on personal data (age, number of times pregnant) and results of medical examination (blood pressure, body mass index, result of glucose tolerance test etc.) . There are eight attributes for each sample. We have divided the entire data set into training and test data consisting of 512 and 256 samples.

Rough set theory is a wonderful approach that offers various solutions for KDD tasks. In order to resolve classification and

description tasks associated with rough set theory, usually the decision rules are considered.

The interesting linguistic features in data mining philosophy are the short and strong decision rules with high confidence which can be formulated in terms of their length, support and confidence. Of the many methods available in RSES for decision rule extraction and decision rule improving the following methods are employed: filtering and shortening of rules.

Table 1 represents reduct generation through Exhaustive algorithm. It is found that no. of reducts is 32. Here the length of the reduct is the number of descriptors in the premise of reducts.

TABLE I RULES THROUGH REDUCT

Algorithm	No. of reducts	Length of Reduct Min	Max	Mean
Exhaustive	32	3	5	3.8

Table II represents a sample of the reducts along with the size, positive region and stability coefficient.

TABLE II REDUCTS THROUGH EXHAUSTIVE

(1-32)	Size	Pos.Reg.	SC	Reducts
1	4	1	1	{PR, PG, DBP, TRICEPS}
2	4	1	1	{PR, PG, DBP, SERUM}
3	3	1	1	{PG, DBP, BMI}
4	3	1	1	{PR, PG, BMI}
5	4	1	1	{PR, PG, DBP, PEDI}
6	4	1	1	{PR, PG, TRICEPS, PEDI}
7	4	1	1	{PR, PG, SERUM, PEDI}
8	3	1	1	{PR, PG, AGE}
9	4	1	1	{PG, DBP, TRICEPS, PEDI}
10	4	1	1	{PR, DBP, TRICEPS, BMI}

Fig 2. and Fig 3 represents the occurences of attributes in reducts and the reduct lengths respectively.

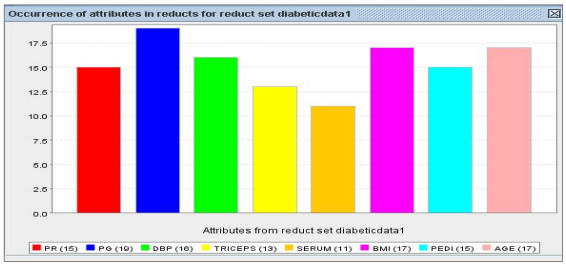

Fig 2. Occurrence of attributes in reducts

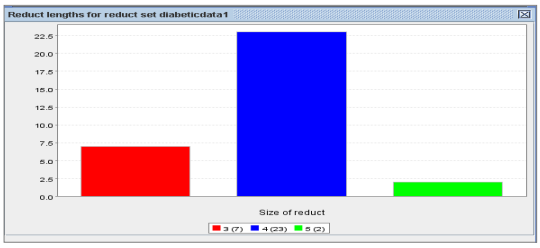

Figure 4. Reduct Length for the reduct set

Table III represents rule generation through reducts by using Exhaustive algorithm. The length of the rule is the number of descriptors in the premise of rules. From the above table, it is found that the accuracy happens to be 71.8%. The rule improvement can be made by filtering the rules.

TABLE III RULE GENERATION

Algorithm	No. of rules	Length of Rules Mini	Max	Mean	Accuracy (%)	Coverage	Filtered rules	Length of Rules Min	Max	Mean
Exhaustive	16364	3	5	3.8	71.8	0.152	20	3	5	3.6

Table IV represents the rule generation by exhaustive, Covering and LEM2 algorithms. In the case of exhaustive the number of filtered rules is 855 and the corresponding accuracy being 67.2%. However, in the case of LEM2, the number of filtered rules is only 114, in which case the accuracy is 76%. The maximum coverage is in the case of exhaustive algorithm. From Tables III & IV, it is clear that reducts certainly helps in improving the rule generation(Case: Exhaustive search).

TABLE IV RULE GENERATION WITHOUT REDUCTS

Algorithms	Rules	Filtered rules	Accuracy (%)	Coverage
Exhaustive	5861	855	67.2	1
Covering	357	150	64.4	0.734
Lem2	300	114	76	0.293

Table V provides the statistics about the decision rules generated from the rough set approach.

TABLE V RULE LENGTH FOR THE INDUCTION ALGORITHMS

Algorithms	Length of rules Min	Max	Mean	Rule Improvement Length of rules Min	Max	Mean
Exhaustive	1	4	2.1	1	3	1.9
Covering	1	1	1	1	1	1
Lem2	2	6	3.5	2	5	3

TABLE VI RULE GENERATION - EXHAUSTIVE RULES

(1-855)	Match	Decision rules
1	14	(PR=1)&(AGE=21)=>(CLASS={tested_negative[14]})
2	13	(PR=2)&(AGE=22)=>(CLASS={tested_negative[13]})
3	12	(PG=99)=>(CLASS={tested_negative[12]})
4	9	(PR=1)&(PEDI=0.1)=>(CLASS={tested_negative[9]})
5	9	(PR=1)&(PEDI=0.5)=>(CLASS={tested_negative[9]})
6	9	(DBP=54)=>(CLASS={tested_negative[9]})
7	8	(DBP=56)=>(CLASS={tested_negative[8]})
8	8	(PEDI=0.4)&(AGE=21)=>(CLASS={tested_negative[8]})
9	8	(DBP=64)&(AGE=21)=>(CLASS={tested_negative[8]})
10	8	(PEDI=0.2)&(AGE=21)=>(CLASS={tested_negative[8]})

Tables VI, VII and VIII provide the details with regard to the rules generated by the three algorithms namely viz., Exhaustive, LEM2 and Covering algorithms. In all the three tables a sample of the rules is presented.

TABLE VIII RULE GENERATION – COVERING RULES

(1-150)	Match	Decision rules
1	12	(PG=99)=>(CLASS={tested_negative[12]})
2	9	(DBP=54)=>(CLASS={tested_negative[9]})
3	8	(DBP=56)=>(CLASS={tested_negative[8]})
4	7	(PG=91)=>(CLASS={tested_negative[7]})
5	7	(PG=94)=>(CLASS={tested_negative[7]})
6	7	(TRICEPS=13)=>(CLASS={tested_negative[7]})
7	7	(TRICEPS=15)=>(CLASS={tested_negative[7]})
8	7	(BMI=27.8)=>(CLASS={tested_negative[7]})
9	6	(PG=92)=>(CLASS={tested_negative[6]})
10	6	(PG=137)=>(CLASS={tested_negative[6]})
11	6	(TRICEPS=21)=>(CLASS={tested_negative[6]})

TABLE VII RULE GENERATION - LEM2 RULES

(1-114)	Mat...	Decision rules
1	13	(PR=2)&(AGE=22)=>(CLASS={tested_negative[13]})
2	9	(PR=1)&(PEDI=0.5)=>(CLASS={tested_negative[9]})
3	9	(PR=1)&(PEDI=0.1)=>(CLASS={tested_negative[9]})
4	6	(SERUM=0)&(PEDI=0.2)&(DBP=78)=>(CLASS={tested_negative[6]})
5	6	(PEDI=0.2)&(AGE=24)=>(CLASS={tested_negative[6]})
6	5	(SERUM=0)&(TRICEPS=0)&(PEDI=0.1)&(PR=2)=>(CLASS={tested_negative[5...
7	5	(PEDI=0.4)&(DBP=60)=>(CLASS={tested_negative[5]})
8	5	(PEDI=0.1)&(AGE=28)=>(CLASS={tested_negative[5]})
9	4	(SERUM=0)&(TRICEPS=0)&(PEDI=0.2)&(PR=3)=>(CLASS={tested_negative[4...
10	4	(SERUM=0)&(TRICEPS=0)&(PEDI=0.2)&(DBP=68)=>(CLASS={tested_negativ...
11	4	(SERUM=0)&(TRICEPS=0)&(PEDI=0.2)&(DBP=70)=>(CLASS={tested_negativ...

V. CONCLUSION

In the performance of data mining and knowledge discovery activities, rough set theory has been regarded as a powerful, feasible and effective methodology since it's very inception in 1982. Generally, medical data contains irrelevant features, uncertainties and missing values. Accordingly, the analysis of such medical data deals with incomplete and inconsistent information with the tremendous manipulation at different levels. In this context, it is emphasized that rough set rule induction algorithms are capable of generating decision rules which can potentially provide new medical insight and profound medical knowledge. By taking into consideration all the above aspects, the present investigation is carried out on the Pima data set consisting of 768 data samples with 8 attributes each. The decision rules are generated for the cases (i) with reducts and (ii) without reducts by using the three rule induction algorithms namely viz., Exhaustive, Covering and LEM2. Our results emphasize that (i) In the case of exhaustive the number of filtered rules is 855 and the corresponding accuracy being 67.2% (ii) In the case of LEM2, the number of filtered rules is only 114, in which case the accuracy is 76% (iii) The maximum coverage is in the case of exhaustive algorithm and (iv) The reducts certainly help in improving the rule generation. Finally, it is concluded that the present results not only provide new medical insight but also are useful for medical experts to analyze the problem effectively.

REFERENCES

[1] Jiawei Han and Micheline Kamber, "Data Mining Concepts and Techniques", San Francisco, CA: Elsevier Inc., 2007.

[2] Piate, U. M. 2006. Tsky-Shapiro, G. & Smyth, P. & Uthurusamy, R. Fayyd, "From Data Mining to Knowledge Discovery: An Overview," in Advances in Knowledge Discovery and Data Mining, 1996a, pp.1-36.

[3] Guoyin Wang , Extension of Rough Set under Incomplete Information Systems ,National Science Foundation of china (No. 69803014), PD program of P.R. China.

[4] Lavrajc, N., E. Keravnou and B. Zupan,." Intelligent Data Analysis in Medicine and Pharmacology",Kluwer Academic Publishers, 1997.

[5] Wolf, S., H. Oliver, S. Herbert and M. Michael,"Intelligent data mining for medical quality management", In Proceedings of the Fifth International Workshop on Intelligent Data Analysis in Medicine and Pharmacology ,Berlin, Germany, 2000.

[6] Jin-Cherng Lin and Kuo-Chiang Wu, "Using Rough Set and Fuzzy Method to Discover the Effects of Acid Rain on the Plant Growth", JCIT, Vol. 2, No. 1, pp. pp ~ 48, 2007.

[7] Ye C.Z., Yang J., Geng D.Y., Zhou Y., Chen N.Y., "Fuzzy rules to predict degree of malignancy in brain glioma", Med. Biol. Comput. Eng. 40 (2), 145~152, 2002.

[8] Lin T.Y., "From rough sets and neighborhood systems to information granulation and computing in words", Proceedings of European Congress on Intelligent Techniques and Soft Computing, 1602~1607, 1997.

[9] Lin T.Y., Yao Y.Y., Zadeh L.A., (Eds.) " Rough Sets, Granular Computing and Data Mining", Physica-Verlag, Heidelberg, 2002.

[10] Tsumoto S., "Mining diagnostic rules from clinical databases using rough sets and medical diagnostic model", Inform. Sci. 162, 65~80, 2004.

[11] Komorowski J., Ohrn A., "Modelling prognostic power of cardiac tests using rough sets", Artif. Intell. Med. 15, 167~191, 1999.

[12] Bazan, J., A. Skowron and P. Synak (1994). Dynamic reducts as a tool for extracting laws from decision tables.In Proc. of the Symp. on Methodologies for Intelligent Systems, Charlotte, NC, October 16–19. Lecture Notesin Artificial Intelligence, vol. 869. Springer– Verlag, Berlin. pp. 346–355.

[13] Hu K.Y., Lu Y.C., Shi C.Y., "Feature ranking in rough sets", AI Commun. 16 (1), 41~50, 2003.

[14] Shifei Ding, Yu Zhang, Li Xu, Jun Qian, "A Feature Selection Algorithm Based on Tolerant Granule", JCIT, Vol. 6, No. 1, pp. 191~ 195, 2011.

[15] Sriman.P.K. and Manjula Sanjay Koti, " A Comparision of different learning models used in data mining for medical datal", The Smithsonian Astrophysics Data System, AIP Conf. Proceedings 1414, 51-55; doi: 10.1063/1.3669930, 2011.

[16] Srimani P. K. and Manjula Sanjay Koti , Cost Sensitivity analysis and prediction of optimal rules for medical data, Proceedings of World Academy of Science, Engineering and Technology 61, DUBAI, pp 1641-1647, 2012.