Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 3, Issue, 12, pp.175-178, December, 2011

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

ISSN: 0975-833X

RESEARCH ARTICLE

THE IMPACT OF ROUGH SET APPROACH ON MEDICAL DIAGNOSIS FOR COST EFFECTIVE FEATURE SELECTION

Prof. Dr. P. K. Srimani[1], F.N.A.Sc., and Manjula Sanjay Koti[2]

[1]Former Chairman, Dept. of Computer Science & Maths, Bangalore University, Director, R&D, B.U., Bangalore
[2]Assistant Professor, Dept. of MCA, Dayananda Sagar College of Engineering, Bangalore
Research Scholar, Bharathiar University, Coimbatore

ARTICLE INFO

Article History:
Received 15th August, 2011
Received in revised form
17th October, 2011
Accepted 27th November, 2011
Published online 31th December, 2011

Key words:
Rough Set (RS), Feature Selection, Rule Induction, Dimension reduction, LEM2, Genetic Algorithm, Reduct, Rules, Cost effectiveness.

ABSTRACT

The medical diagnosis process can be interpreted as a decision making process, during which the physician includes the diagnosis of a new and unknown case from an available set of clinical data and from clinical experience which can be computerized. A method that enhances the performance of a model that is based on Rough set theory for feature selection and classification is proposed. For this purpose, the PIMA dataset is used. The proposed system provides the solution to a feature subset selection problem which is nothing but a task of identifying and analysing an optimal subset from a larger set of features. It is concluded that the method certainly helps in cost reduction associated with the diagnosis.

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 (Han and Kamber,2007). The process of finding useful patterns or meaning in raw data has been called knowledge discovery in databases (Piate *et al.,*1996). 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). In fact, the growth in the size of data and the number of existing databases far exceed the ability of humans to analyse this data, which creates both a need and an opportunity to extract knowledge from databases (Cios *et al.,*1998). 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.

***Corresponding author:** profsrimanipk@gmail.com, man2san@rediffmail.com*

The huge amount of data and knowledge stored in medical databases require sophisticated tools for storing, accessing, analysis, effective and efficient usage of stored knowledge and data. Further, Intelligent methods such as neural networks, fuzzy sets, decision trees and expert systems are applied in the medical fields. In recent years, some applications of a new intelligent technique known as Rough set theory has been applied to discover data dependencies, data reduction, approximate set classification and rule induction from huge databases.

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 (mmHg), 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). The theory of rough sets (Skowron *et al.,* 2002) has emerged as a major mathematical tool for managing uncertainty that arises from granularity in the domain of discourse. 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 satisfying classification.
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. The present investigation on Rough sets for cost effectiveness is organised as follows: Related work, Methodology, Experiments and Results. Finally the conclusions are presented.

Related work

There have been many studies applying data mining techniques to the PIMA (PIDD) (Srimani and Manjula, 2011). The independent or target variable is diabetes status. Some of the related works include (Quinlan, 1998)(Smith etal., 1988) (Ephzibah,2011). No work pertaining to the topic of research is available. Hence the recent investigation is carried out.

METHODOLOGY

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 is easy internet access to the rich literature about the rough set theory, its extensions as well as interesting applications.

Dimensionality reduction

Dimensional reduction has been a major factor in data mining problems. In many real time situations, e.g. database applications and bioinformatics, there are far too many attributes to be handled by learning schemes, majority of them being redundant. Taking predominant attributes reduces the dimensions of the data, which in turn reduces the size of the hypothesis space, and thereby allowing classification algorithm to operate faster and more efficiently. The Rough Set (RS) theory is one such approach for dimension reduction. RS offers a simplified search for predominant attributes in datasets.

Information Systems

A data set is represented as a table, where each row represents a case, an event, a patient or simply an object(Grzymala-Busse, 2004). Every column represents an attribute that can be measured for each object; the attribute may be also supplied by a human expert or the user. This is represented as a pair $S = (U, A)$ where U is a non-empty finite set of *objects* called the *universe* and A is a non-empty finite set of *attributes* such that $a : U \rightarrow V_a$ for every $a \in A$. The set V_a is called the *value set* of a.

Reduct

A reduct is a set of necessary minimum data, since the original proprieties of the system or information table are maintained. Therefore, the reduct must have the capacity to classify objects, without altering the form of representing the knowledge. Reduct and core of condition attributes helps in removing of superfluous partitions (equivalence relations) or/and superfluous basic categories in the knowledge base in such a way that the set of elementary categories in the knowledge base is preserved. This procedure enables us to eliminate all unnecessary knowledge from the knowledge base and preserving only that part of the knowledge which is really useful.

Rule Induction

Rule induction is one of the most important techniques of machine learning. Regularities hidden in data are frequently expressed in terms of rules; rule induction is one of the fundamental tools of data mining. Rules are generally in the following form: If (attribute1, value1) and (attribute2, value2) and (attribute, value) then (decision, value) Data from which rules are induced are usually presented in a form similar to a table in which cases (or examples) are labels (or names) for rows and variables are labelled as attributes and a decision. Attributes are independent variables and the decision is a dependent variable. The set of all cases labelled by the same decision value is called a concept.

We have used the following algorithms in our investigation:
Exhaustive search algorithm starts with an empty feature set and carries out exhaustive search until it finds a minimal combination of features that are sufficient for the data analysis task which works on binary, noise-free data and runs in the time of $O(N^M)$, where N is no. of tuples, and M is the number of attributes. Sequential covering algorithm sequentially learns a set of rules that together cover the whole set of positive examples. Genetic algorithms are search algorithms based on the mechanics of natural selection and natural genetics . LEM2 (Grzymala-Busse, 1997) explores the search space of attribute-value pairs.
The procedure LEM2 is presented below.
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;
$C = C \cup \{c\}; G = [c] \cap G;$
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;
$G == B - U \quad |R|$
$r \in R$
end;
for each r ∈R do
if U |S|=B then R = R –r
$s \in R - r$
end

Fig1. LEM2 Algorithm

The Rough set philosophy was founded on assumption that every object of the universe set associated with some information (knowledge, data) (Skowron *et al.,* 2002). 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 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.

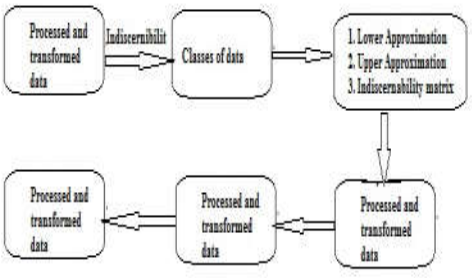

Fig 2. Mining the patient data using rough set

EXPERIMENTS AND RESULTS

We have used 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 500 samples of class 1 and 268 of class 2. There are eight attributes for each sample. We have used reducts and rules are generated. The results of the present investigation are presented in Tables 1,2, 3 and 4; Fig's. 2,3,4, 5 and 6.

Table 1. Set of reducts

Size	Positive region	Stability coefficient	Reducts
3	1	1	{PG,DBP,BMI}
4	1	1	(PR,PG,SERUM,AGE}
4	1	1	{PR,PG,SERUM,BMI}
4	1	1	{PR,PG,TRICEPS,PEDI}
4	1	1	{PG,PG,TRICEPS,AGE}
4	1	1	{PR,,PG,BMI,PEDI}
4	1	1	{PR,TRICEPS,BMI,PEDI}
3	1	1	{PG,BMI,AGE}
4	1	1	(PG,SERUM,PEDI,AGE}
4	1	1	{PG,TRICEPS,PEDI,AGE}

Rough set data analysis was applied to the PIMA data to find the reducts and core of the data. We have found that the PIMA data set of 768 patients have ten reducts. Table 1 represents a sample of result reducts which concentrate with no more than four elements. But the core of the system is empty. This signifies a huge inhomogeneity among the attributes. In other words the dependency among the attributes is high, and there are many possibilities for substitution.

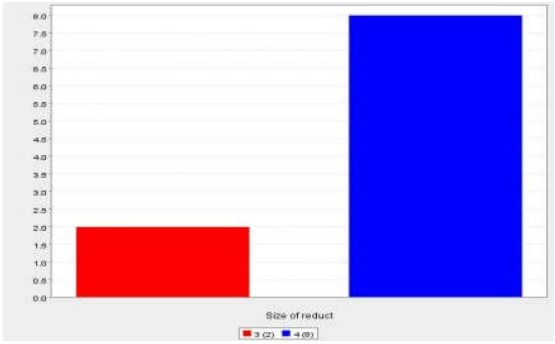

Fig. 3. Reduct length for reduct se
t

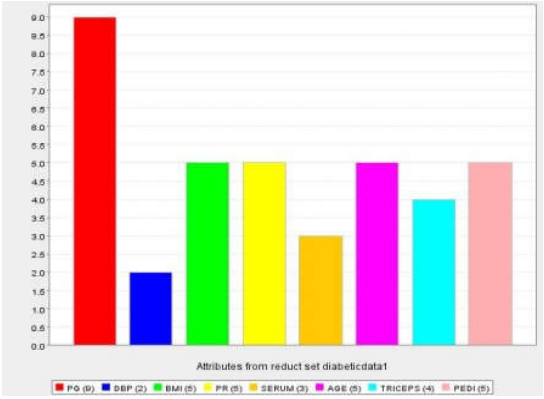

Fig.4. Occurrence of attributes in reducts

From Table 4. it is found that the present approach is much more effective when compared to the available results. The accuracy obtained in the present work is 100%, while that in the available work is 87%. The cost reduction obtained through the feature selection procedure based on rough set approach is found to be optimal in the present work.

Table 2. Rules generated (partialset) for lem2 algorithm

Rules	Instances
IF (PR=2) & (AGE=22) THEN CLASS=NO	18
IF (SERUM=0)&(PR=1)&(PEDI=0.1)THEN CLASS=NO	10
IF(SERUM=0)&(TRICEPS=0)&(PEDI=0.2)&(PR=3) THEN CLASS=NO	5
IF (SERUM=0)&(PEDI=0.3)&(DBP=76)THEN CLASS=YES	4
IF(SERUM=0)&(TRICEPS=0)&(PEDI=0.5)&(PR=4) THEN CLASS=YES	3

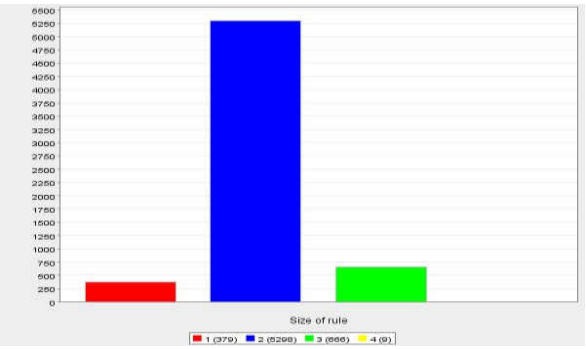

Fig.5. Rule lengths for the rule set

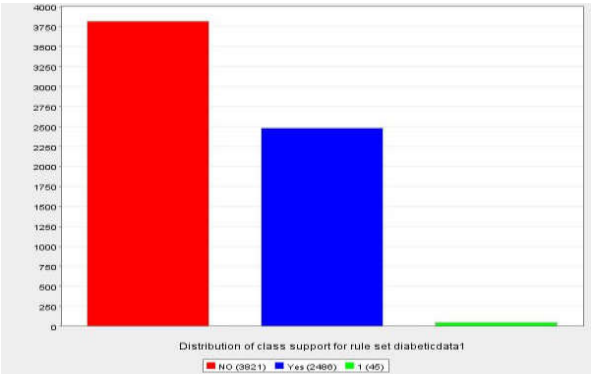

Fig. 6. Number of rules supporting decision classes

Table 3. Accuracy and coverage for rule generation

Algorithms	No. of Rules	Filtered Rules	Accuracy	Coverage
Exhaustive	379	166	100	100
Genetic	6352	1288	100	100
Covering	9351	1417	100	63.4
LEM2	453	166	100	89.8

Table 4. Include % with Accuracy %

		Original features	Reduced feature	Accuracy	Cost
Our Results	Without GA	8	-	65.5%	100
	With GA	8	4	100	50
Ephzibah (2011)	Without GA	8	-	69	100
	With GA	8	5	87	62

CONCLUSION

Generally, people expect an optimal approach for the diagnosis of any disease. Feature selection is a technique that reduces or lessens the number of features. In medical world, for any disease to be diagnosed there are some tests to be performed and each and every test can be considered as a feature. By the process of feature selection, the performance of tests that are highly expensive and irrevalant could be avoided, which in turn reduces the cost associated with the diagnosis and helps the patients and the doctors to a great extent. In processing the medical data, choosing the optimal subset of features is important, not only to reduce the processing cost but also to improve the classification performance of the model built from the selected data. Rough Set method has been recognized to be one of the powerful tools in the medical feature selection. However, the high storage space and the time-consuming computation restrict its application. The present investigation certainly helps in cost reduction associated with the diagnosis, which in turn facilitates the patients and doctors considerably.

ACKNOWLEDGEMENT

One of us Mrs. Manjula Sanjay Koti is grateful to Bharathiar University, Tamil Nadu for providing the facilities to carry out the research work.

REFERENCES

Skowron , A., Z. Pawlak, J. Komorowski and L. Polkowski,: "*A rough set perspective on data and knowledge, in W. Kloesgen, J. Żytkow (Eds.): Handbook of KDD*". Oxford University Press, Oxford (2002), 134-149.

Cios, K.,W. Pedrycz and R. Swiniarski (1998). "*Data Mining Methods for Knowledge Discovery"*. Kluwer Academic Publishers.

Srimani, P. K. and Manjula Sanjay Koti, "*„Application Of Data Mining Techniques For Outlier Mining In Medical Databases", IJCR,, Vol. 33, Issue, 6, pp.402-407, June, 2011.*

Srimani, P. K. and Manjula Sanjay Koti, "*A comparison of different learning models used in data mining for medical data*", In Press.

Ephzibah, E.P. 2011. "*Cost effective approach on feature Selection using genetic algorithms and Fuzzy logic for diabetes diagnosis."* IJSC.

Grzymala-Busse, J. W. 2004. "*Three Approaches to Missing Attribute Values - A Rough Set Perspective"*, Workshop on Foundations of Data Mining, associated with the fourth IEEE International Conference on Data Mining, Brighton, UK, November 1–4.

Grzymala-Busse, J.W.1988. "*Knowledge acquisition under uncertainty - A rough set approach"*. Journal of Intelligent & Robotic Systems", 1: 3{16).

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

Quinlan, J. R. 1993. "C4.5: *Programs for Machine Learning*". San Mateo, Calif., Morgan Kaufmann Publishers.

Smith, J. W., et al.1988. "*Using the ADAP learning algorithm to forecast the onset of diabetes mellitus". Proceedings of the Symposium on Computer Applications and Medical Care* (Washington, DC). R. A. Greenes. Los Angeles, CA, IEEE Computer Society Press**:** 261-265.

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.