ROUGH SET APPROACH FOR NOVEL DECISION MAKING IN MEDICAL DATA FOR RULE GENERATION AND COST SENSITIVENESS

N Divya
Research scholar, Department of CSE, SSUTMS- Sehore, Madhya Pradesh, India,
naademdivya@gmail.com

VVSSS Balaram
Professor, HOD Department of Information Technology, Sreenidhi Institute of Science And Technology, Hyderabad, Telangana, India.

Abstract:
Clinical decision support systems (CDSS) often base on rules that are inferred from collected patients' histories, together with expert judgements and consented medical guidelines. This type of advisor system is known as rule-based reasoning system or expert system which classifies a given test instance into a particular outcome from the learned rules. The results clearly show that rough set approach is certainly a useful tool for medical applications. Relationships and patterns within this data could provide new medical knowledge. The genetic algorithms offer an attractive approach for solving the feature subset selection problem. The process of finding useful patterns or meaning in raw data has been called knowledge discovery in databases. The algorithms used for the present study are: Exhaustive search, Covering, and Genetic algorithms

Keywords*: medical applications, Exhaustive search, Covering, and Genetic algorithms*

I.INTRODUCTION

Data mining strategies can be applied in the zone of Software Engineering for getting improved outcomes. Medical data mining has incredible potential for exploring the shrouded patterns in medical data and these patterns can be used for clinical finding. Examination of medical data is often concerned with the treatment of incomplete knowledge, with management of inconsistent snippets of information. In the present investigation, the theory of rough set is applied to find reliance relationship among data, assess the importance of qualities, find the patterns of data, learn common decision-making rules, lessen the redundancies and look for the minimum subset of ascribes to attain satisfactory classification. It is concluded that the decision rules (with and without reducts) created by the harsh set induction algorithms (Exhaustive, Covering, GA) give new medical insight as well as are valuable for medical specialists to break down the issue viably and find ideal expense.

Overview of Rough Set Theory (RST):

In recent years, Classical rough set theory developed has made a great success in knowledge acquisition. Rough set theory is relatively a new tool that deals with vagueness and uncertainty inherent in decision making. Rough Set Theory (RST) is a useful mathematical tool to deal with imprecise and insufficient knowledge, find hidden patterns in data, and reduce the size of the dataset. Also, it facilitates (i) the evaluation of the significance of the data and (ii) the easy interpretation of the results. Rough set theory has been regarded since its very inception as a powerful and feasible methodology for performing data mining and knowledge discovery activities set theory, knowledge is represented in information systems and an information system is a data set represented in the form of a table, which is called as decision table. In the area of knowledge discovery particularly in machine learning and rule extraction,

rough set theory has gained much popularity. In particular, it is useful in discarding redundant information in a database, i.e. to arrive at a minimum number of attributes that would still allow each data record to be distinguished from the others. This minimum set of attributes is called as a reduct.

II.LITERATURE REVIEW

In the past computers were used in health care mainly for administrative tasks and medical informatics. A new trend of interactive computer applications emerged called clinical decision support system (CDSS) Shu-Li Wang [1] that assists clinicians with decision making tasks. Central to CDSS is the knowledge which is accumulated from medical data collected from patients. Though the process of gathering the data might already be in place, especially for those health care systems that digitize and archive medical records electronically, how CDSS can manipulate these data and assist in the decision making remains a challenge Wang W [2] These approaches include practicing evidence-based medicines Randolph AG [3], enforcing health-care protocols by Morris AH [4] and proposing expert guidelines for clinical practice. For instance, a health institute can initiate a consensus procedure; start with tapping on the intangible knowledge and experiences from the medical experts for a draft, then refining the draft into a form of decision rules through clinical experiences and opinions, and finally getting the experts to consent them into guidelines. Standard guidelines that can be readily installed into computerized CDSS rarely exist by Reed M Gardner [5]. Hybrid solutions are usually tightly coupled where two or more models are fused together, mathematically they compute a quality outcome; the downside however is the lack of human understandable rules. He, J., Hu [6]. In addition to hybrid models, ensemble model is recently regarded as a popular approach in providing a collective prediction result Eom; Jae-Hong [7]. Ensemble or ensemble learning differs mainly from hybrid methods in its loosely coupling the learners. A pool of individual learners operates independently, sometimes with different hypotheses, and at the end a winner learner is chosen for its better predictive performance. Kamiran et al. [8] pointed out that many of the methods that make classifiers aware of discriminatory biases require data modifications or algorithm tweaks and they are not very flexible with respect to multiple sensitive feature handling and control over the performance vs. discrimination trade-off. Liu et al. [9] highlighted that most work in machine learning fairness had mostly studied the notion of fairness within static environments, and it had not been concerned with how decisions change the underlying population over time. They argued that seemingly fair decision rules have the potential to cause harm to disadvantaged groups and presented the case of loan decisions as an example where the introduction of seemingly fair rules can all decrease the credit score of the affected population over time. Milli et al. [10] also studied how individuals adjust their behaviour strategically to manipulate decision rules in order to gain favourable treatment from decision-making models. They reiterated that the design of more conservative decision boundaries in an effort to enhance robustness of decision-making systems against such forms of distributional shift is significantly needed in order for fairness to be achieved.

III. RESEARCH METHODOLOGY

The present examination outlines how harsh set theory could be utilized for the investigation of medical data particularly for generating classification rules from a lot of watched tests of the Pima data set. In unpleasant set theory, knowledge is spoken to in information systems. An information system is a data set spoken to in a forbidden form considered decision table in which each line speaks to an item and every section speaks to a property. In order to determine all the reducts of the data that contains the minimal subset of characteristics that are related with a class name for classification. The Rough Set reduction system is utilized. In a knowledge system reducts are often utilized at the data preprocessing stage during the quality selection process. It is important to take note of that reduct isn't exceptional and in a decision table different reducts may exists. The core of a decision table which consists of basic information is certainly contained in each reduct. In other words, a reduct produced from the original data set ought to contain the core properties. During the trait selection process reduct and core are the commonly utilized since the main reason for unpleasant set theory is to choose the most important ascribes with respect to the classification task and to evacuate the unimportant qualities. The arrangement of credits which is common to all reducts is known as the core which is controlled by each authentic reduct, and subsequently consists of basic qualities which can't be expelled from the information system without causing breakdown of the proportionality class structure. In other words, a core is completely essential for the representation of the categorical structure.

Rule Improvement by RSES:

After calculation of decision rules, one can select some most interesting rules for further usage. Some time we cannot do it because of the low support of calculated rules. In RSES, the support of decision rules can be increased by the following ways.

Rule generation → (Discretization (local or global) + Algorithms + Rule shortening).

Discretization of real value attributes: two discretization methods called "local" and "global" methods have been implemented in RSES. Both the methods are based on Boolean reasoning approach. The local method is built on decision tree and it usually products more cuts than global method.

Rule shortening: Removing some descriptors from a given decision rule can increase its support, but it decreases its confidence. In RSES, we can determine the "shortening ratio", which is a minimal acceptable threshold for confidences of decision rules in shortening process.

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:

Genetic Algorithm

One can compute a predefined number of minimal consistent rules with genetic algorithm that comprises permutation encoding and special crossover operator.

1	Choose initial population
2	Evaluate the fitness of each individual in the population
3	
4	repeat

5	Select best ranking individuals to reproduce
6	
7	Breed new generation through crossover and mutation Genetic operations) and give birth to offspring
8	Evaluate the individual fitness's of the offspring Replaced the worst ranked part of population with offspring Until < terminating condition>

Figure 3.1: Genetic Algorithm

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.

1	Exhaustive (intsol,intdepth)
2	{if(issolution)(sol)
3	Printsolution(sol)
4	else
5	{solgenarated=genaratedsolution()
6	Exhaustive (sol generated, depth+1)}

Figure 3.2: Exhaustive Algorithm

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

1	Input: labeled training dataset D
2	Outputs: ruleset R that covers all instances in D
3	
4	Procedure
5	Initialize R as the empty set
6	For each class C {
7	While D is nonempty {
8	Construct one rule r that correctly
9	classifiers some instances in D that
10	belong to class C and does not
11	Incorrectly classify any non -C instances
12	Add rule r to ruleset R Remove from D all instances correctly classified by r}} return r

Figure 3.3: Covering Algorithm

Back Propagation Network: BP algorithm is a supervised learning method, which it is the most widely used algorithm for training MLP neural network. The idea of the BP is to reduce

this error, until the ANN learns the training data. The training begins with random weights, and the goal is to adjust them so that the learning error will be at minimal. ANN nodes in BP algorithm are organized in layers, send their signals forward and then the learning error (difference between actual and expected results) is calculated and propagated backwards until met satisfactory learning error.

the interconnection between nodes which is usually referred to as a fully connected network or multilayer perceptron (MLP). Multilayer architecture means that the network has several layers or nodes usually referred to as input layer, hidden layer and output layer. MLP network can be used with great success to solve both classification and function approximation problems.

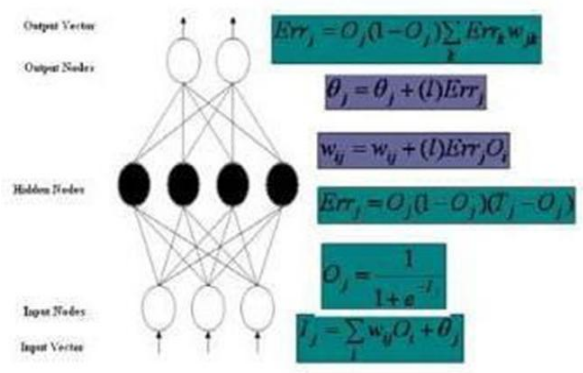

Figure 3.4: Multi-Layer Perceptron Structure

IV. EXPERIMENTS AND RESULTS

The results of the experiments conducted on the Pima data set by using rough set approach are presented and discussed in this section

Table 4.1: Rules Through Reduct for Pima Data Set

Algorithm	No. of reducts	Length of Reduct Min	Max	Mean
Exhaustive	32	3	5	3.8
Genetic	10	3	4	3.4

Table 4.2: Rule Generation for PIMA Data Set

Algorithm	No. of rules	Length of Rules Min	Max	Mean	Accuracy (%)	Coverage	Filtered rules	Length of Rules Min	Max	Mean
Exhaustive	16364	3	5	3.8	71.82	0.152	2010	3	5	3.6
Genetic	511	3	4	3.4	78.10	0.1	10	3	3.	3.
tic	0			6			6	1		

Table 4.1 represents reduct generation through Exhaustive and Genetic algorithms. It is discovered that the quantity of reducts is more for the situation of comprehensive. Here the length of the reduct is the quantity of descriptors in the reason of reducts. Table 4.2 represents rule generation through reducts by using these algorithms and likewise the classification results. The length of the standard is the quantity of descriptors in the reason of guidelines. Here, the precision happens to be more for the situation of genetic algorithm despite the fact that the inclusion is less when contrasted with the comprehensive algorithm. Table 4.3 presents the length of the guidelines.

Table 4.3: Rule Generation-Direct for PIMA Data Set

Algorithms	Rules	Filtered rules	Accuracy (%)	Coverage	Std. dev.
Exhaustive	5861	855	67.2	1	-
Covering	357	150	64.4	0.734	-
Lem2	300	114	76	0.293	-
Genetic	3574	749	64.26	0.990	3.14

For the situation of reducts generation based on genetic algorithms actualized through RSES, it is discovered that the outcomes will be non-deterministic and this is certainly the consequence of a genetic algorithm worldview. Accordingly various correctness's for various executions will be obtained for similar data

sets. Therefore, it is completely important to explore commonly and process the normal precision and the standard deviation. Further, it is proposed that only those outcomes for which little estimations of standard deviation are obtained would be worthy. A similar procedure is followed in Tables 4.4-4.5. No place in the writing, results corresponding to the genetic algorithm executed through RSES is accessible.

Table 4.4: Rule Generation-Direct for Pima Data Set (Length of Rules)

Algorithms	Length of Rules		
	Min.	Max.	Mean
Exhaustive	1	4	2.1
Covering	1	1	1
Lem2	2	6	3.5
Genetic	1	4	2

In medical world, for any malady to be analyzed there are a few tests to be performed and every single test can be considered as a component. By the procedure of highlight selection, the performance of tests that are profoundly costly and unimportant could be dodged, which in turn diminishes the expense related with the determination and helps the patients and the doctors as it were. In processing the medical data, choosing the ideal subset of highlights is important, not exclusively to diminish the processing cost yet additionally to improve the classification performance of the model worked from the chose data.

Table 4.5 predicts the following: with GA (i) the number of features reduced is 4 ii) accuracy is 74.8% (iii) Time is 0.02 ms (iv) ROC value being 79.1% and (v) optimal cost is 25%.

Table 4.5: Optimal Cost Prediction for PIMA Data Set

Sl. No	Algorithm	Original features	Features reduced	Accuracy (%)	ROC (%)	Cost (units)	Average Cost (%)	Time (ms)
1	With GA	8	4	74.8	79.1	193	25%	0.02
2	Without GA	8	-	73.8	75.1	201	26%	0.06

Evaluating The Performance of The Proposed Algorithm

In order to show the advantages of our Distributed Association Rule Mining for Predicting Heart Diseases algorithm (DRAM-PHD), we have performed a number of tests that demonstrate the ability of the proposed algorithm to work correctly in a distributed knowledge environment without moving all the databases to a single site. The tests were performed to find out the effect of various parameters on the final result. The two important variables that affect the result are: 1) the number of sites. 2) the number of tuples per database. These tests have been carried out on a network of workstations connected by a LAN and tested against a number of databases of different sizes. We compare our results with the results of the algorithm in where the authors solved the same problem in vertically distributed databases which will be same as horizontally distributed if all attributes are shared among the sites, i.e., all attributes are the same in all local databases

Changing Number of Sites: The first test was executed to demonstrate how the elapsed time and the number of exchanged messages varies with the number.

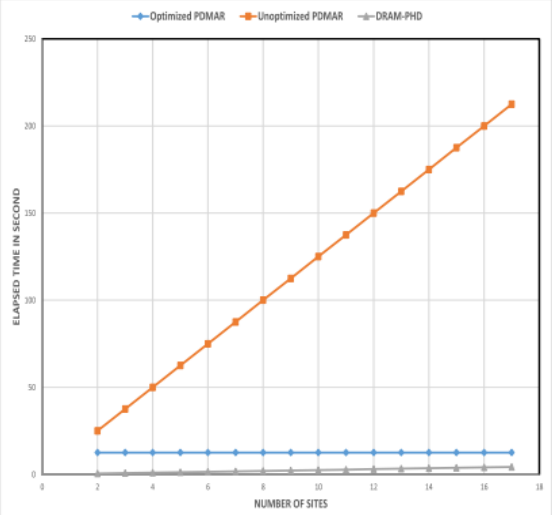

Figure 4.1: Optimized PDMAR, Un-optimized PDMAR and the proposed algorithm.

Figure 4.1 The number of exchanged messages between the participating sites and the coordinator site in the proposed algorithm, Optimized PDMAR and Un-optimized PDMAR. It can be seen easily that the number of exchanged messages increases as the number of local sites increases and the proposed algorithm has the smallest number of exchanged messages. This because the Optimized PDMAR and the Un-optimized PDMAR algorithms create a relation called shared and deal with each set of tuples that correspondence to a shared tuple as a class.

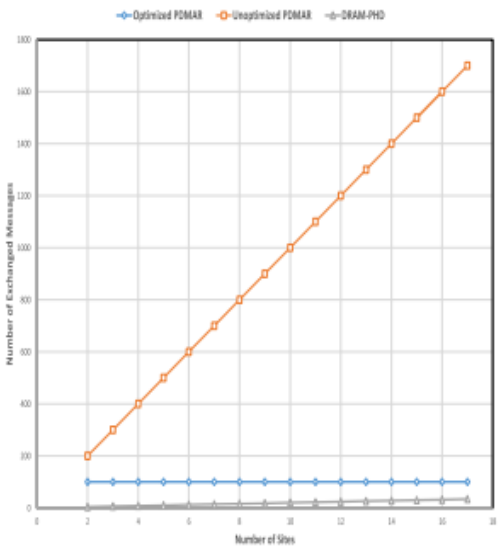

Figure 4.2: Elapsed time versus number of local sites in Optimized PDMAR, Un-optimized PDMAR and the proposed algorithm.

The elapsed time to find the association rules in an implicit database D changes with the number of local sites in the proposed, Optimized PDMAR and Un-optimized PDMAR algorithms. It can be seen easily that the elapsed time increases as the number of local sites increases. and the proposed algorithm has the least elapsed time because the elapsed time depends on the number of exchanged messages

Discussions:

Healthcare around the world is committed to providing quality care to patients via electronic health records. Due to the distributed nature of the EHRs, shared access to health records should be made possible and data integration should be established.

- Preserving the privacy of patient information is an important consideration when handling medical data to developed a privacy-preserving integration model based on association rules for predicting heart disease using

patient data collected from horizontally distributed databases.

- Our model allows the sharing of data summaries (useful information) to be used to predict heart disease.
- These summaries are not accompanied by private patient information. Our approach is the first to use association rules metrics for naturally distributed medical datasets to generate weighted rules, which are further generalized using independent test datasets rather than using specific rules for each local model. In the future, work to discover more privacy-preserving integration models for vertically and horizontally distributed systems can be considered.

Conclusion

In the context of numerous down to earth issues like medical finding, highlight subset selection introduces a multi-criterion (viz., exactness of classification, cost and hazard related with classification) optimization issue which encourages the meaningful pattern recognition. These criteria strongly rely upon the selection of traits that depict the patterns. Truth be told, the main thought of highlight selection is to choose a subset of input factors by eliminating the component/traits with next to zero prescient information. For the most part, medical data contains immaterial highlights, uncertainties and missing qualities. Accordingly, the examination of such the consequences of the present investigation strongly uncovers the following:

- The quantity of reducts is more for the situation of thorough algorithm
- the exactness happens to be more for the situation of genetic algorithm in spite of the fact that the inclusion is less when contrasted with the comprehensive algorithm
- the quantity of highlights decreased is 4, exactness is 74.8%, Time is 0.02 ms, ROC worth being 79.1% and ideal expense is 25% with and without GA.
- Finally, it is concluded that the outcomes obtained from GA executed through RSES is of non-deterministic nature and consequently exact outcomes can be obtained only by computing the normal of a few keeps running for similar data set. This is first of its kind and not found in the writing.

References:

1. *Shu-Li Wang, Shih-Yi Yeh, Framework of Computer-Assisted Instruction and Clinical Decision Support System for Orthodontics with Case-Based Reasoning, Proc. International Conference on Medical Biometrics, Lecture Notes in Computer Science, 2010, 344-352.*
2. *Wang W, Richards G, Rea S., Hybrid data mining ensemble for predicting osteoporosis risk, Conf Proc IEEE Eng Med Biol Soc. 2005;1:886-9.*
3. *Randolph AG, A practical approach to evidence-based medicine: lessons learned from developing ventilator management protocols. Crit Care Clin, 2003, 19(3), 515–527.*
4. *Morris AH, Treatment algorithms and protocolized care. Curr Opin Crit Care, 2003, 9(3), 236–240.*
5. *Reed M Gardner, Computerized Clinical Decision Support in Respiratory Care, Respiratory Care, April 2004, 49(4), 378-388.*
6. *He, J., Hu, H. J., Harrison, R., Tai, P. C. and Pan, Y., Transmembrane segments prediction and understanding using support vector machine and decision tree.*

Elsevier Expert Systems with Applications Journal, 30(1) (2006), 64–72

7. *Eom; Jae-Hong, Kim; Sung-Chun and Zhang; Byoung-Tak, Apta CDSS-E: A classifier ensemble-based clinical decision support system for cardiovascular disease level prediction, Elsevier Expert Systems with Applications Journal, Volume 34 (2008), 2465–2479*
8. *Kamiran, F.; Karim, A.; Zhang, X. Decision theory for discrimination-aware classification. In Proceedings of the 2012 IEEE 12th International Conference on Data Mining, Brussels, Belgium, 10–13 December 2012; pp. 924–929.*
9. *Liu, L.T.; Dean, S.; Rolf, E.; Simchowitz, M.; Hardt, M. Delayed Impact of Fair Machine Learning. In Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, 10–15 July 2018; Volume 80, pp. 3156–3164*
10. *Milli, S.; Miller, J.; Dragan, A.D.; Hardt, M. The social cost of strategic classification. In Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, 29–31 January 2019; pp. 230–239.*