AENSI Journals

Australian Journal of Basic and Applied Sciences

ISSN:1991-8178

Journal home page: www.ajbasweb.com

Feature Granularity of Cardiac Dataset using Rough Set Technique

[1]S.N. Suhana and [2] S.S.Mariyam

[1]*Faculty of Computer, Media and Technology Management,TATI University College Jalan Panchor, Teluk Kalong, 24000 Kemaman, Terengganu.*
[2] *Soft Computing Research Group, Universiti Teknologi Malaysia,1310 Skudai, Johor, Malaysia.*

ARTICLE INFO

Article history:
Received 20 November 2013
Received in revised form 24 January 2014
Accepted 29 January 2014
Available online 5 April 2014

Key words:
Rough Set, Reducts, Rules, Classification, Medical Data

ABSTRACT

Rough Set is a remarkable technique that has been successfully implemented in diverse applications including medical field. Typically, Rough Set is an efficient instrument in dealing with huge dataset in concert with missing values and granularing the features. However, large numbers of generated features reducts and rules must be chosen cautiously to reduce the processing power in dealing with massive parameters for classification. Hence, the primary objective of this study is to probe the significant reducts and rules prior to classification process of cardiac datasets from National Heart Institute (NHI), Malaysia. All-embracing analyses are presented to eradicate the insignificant attributes, reduct and rules for better classification taxonomy. Reducts with core attributes and minimal cardinality are preferred to construct new decision table, and subsequently generate high classification rates. In addition, rules with highest support, fewer length and high Rule Importance Measure (RIM) are favored since they reveal high quality performance. The results are compared in terms of the classification accuracy between the original decision table and a new decision table. It demonstrates that the rules with highest support value are more significant compared to the rules with less length.

To Cite This Article: S.N. Suhana and S.S.Mariyam, Feature Granularity of Cardiac Dataset using Rough Set Technique. ***Aust. J. Basic & Appl. Sci.,*** *8(4): 114-122, 2014*

INTRODUCTION

Medical data such as breast cancer regularly contain irrelevant features as such uncertainties and missing values exist. The analysis of medical data frequently requires dealing with imperfect and conflicting information, superfluous objects and attributes as well as exploitation of various levels of data representation. In the meantime, the effect of soft computing novelty solves more application domains gradually, and principally in medical field. The classification is the most crucial part in persuading the treatment of patients. A compact classification system with minimum number of variables in medical decision making will furnish these data for further analysis in small interval and in an intelligible format. Medical domain often been employed for newly developed learning and reasoning techniques from any techniques (Øhrn, 1999). The medical field has many applications whose solutions are important in a social context. However, the field is notoriously difficult with a wide range of confounding factors and aspects that demand and medical domain has led to some remarkable work being accomplished in the last 30 years, but major developmental efforts and research is still needed if a significant impact on the practice of medicine is to be realized. Medical data is a huge dataset consist of attributes including the missing value inside, possibly have similar redundant information that needs to be discarded. Due to the abundance of noisy, irrelevant or misleading features, the ability to handle imprecise and inconsistent information in medical domain has become one of the most important requirements for feature selection and classification (Wang *et al*., 2007)

FRE-FMMNN was found to produce better predictions than the other methods compared to decision trees (Quinlan, 1986) a multi-layer perceptron (MLP) network trained with a backpropagation (BP) algorithm (Zurada & Jacek, 1992) and a nearest-neighborhood method (Andrew, 1999). However, in the classification problem, FREFMMNN algorithm produce only two rules, which may not be sufficient for medical experts to analyze brain glioma data and find the real cause-and-effect dependency relations between glioma MRI features and the degree of malignancy. The membership functions and sensitivity parameter also must be set beforehand (Ye *et al*., 2002). Support vector machine (SVM) has been found useful in handling classification tasks in case of the high dimensionality and sparsity of data points and has been among as a popular approach to efficiently treating the medical data structure. Although the approach of SVM with kernel function is useful for

Corresponding Author: S.N. Suhana, Faculty of Computer, Media and Technology Management,TATI University College Jalan Panchor, Teluk Kalong, 24000 Kemaman, Terengganu.

classification, however the computation speed is relatively slow when the kernel functions are complicated. Instead, its performance must be improved especially for complex data (Su and Yang, 2008). This is particularly important for people who want to obtain a high level of accuracy in advanced areas such as precision engineering and medical diagnosis. The heuristic search exponential method includes methods such as branch and bound (Narendra and Fukunaga, 1977) which starts from a full set and removes features using a first depth strategy. The method guarantees an optimal solution under the monotonic assumption that the children of the nodes whose objective function values are lesser than the current best not contain a better solution and so these features not further explored.

Rough set is a fairly useful clever technique that has been applied to the medical domain and is used for the discovery of data dependencies, evaluates the importance of attributes, discovers the patterns of data, reduces all redundant objects and attributes, and seeks the minimum subset of attributes. From the medical point of view, this aims at identifying subsets of the most vital attributes influencing the treatment of patients. The chosen subsets are then engaged within a decision rule generation process, creating descriptive rules for the classification task, which may potentially reveal profound medical knowledge and provide new medical insight. These decision rules are more useful for medical experts to analyze and gain understanding into the problem at hand. Decision rules extracted by rough s*et al*gorithms are concise and valuable, which can be benefit to the medical experts by enlightening some knowledge hidden in the data.

In rough information system, there often exist some condition attributes that do not provide any additional information about the objects. Consequently, those attributes should be reduced if those condition attributes are eliminated. A decision table may have more than one reduct. Any decision table can be used to replace the original table. Furthermore, the best reduct with the most minimal number of attributes criteria should be consider if there are two or more reducts with the same number of attributes occurred. If this case happened, then the reducts with the least number of combinations of values attributes is selected. The rules derivations are generated from the reduct. There are possible to have a lot of rules to use in classification process. However, not all the rules are significant for better classification. Hence, significant reducts and rules must be chosen to contribute better classification. Subsequently, core attributes and minimal cardinality are exploited to expedite the significant reducts and the rules measurement of rules support, rules length, rule important measure (RIM). The results show that the proposed method give better classification performance compared to standard rough set in terms of accuracy.

2. Literature Review:

Rough-Set Theory (RST) was introduced by Polish logician, Professor Zdzisław Pawlak (1982) to cope with imprecise or vague concepts. Recently, it is one of the most developing soft computing methods for the identification and recognition of common patterns in data, especially in the case of uncertain and incomplete data. The mathematical foundations of this method are based on the set approximation of the classification space (Pawlak, 1991). A rough set is a formal approximation of a crisp set which is *conventional set*, in terms of a pair of sets which give the *lower and the upper approximation* of the original set (Pawlak, 1998).

Knowledge base for rough set processing is stored as a table containing conditional and decision attributes. A method of knowledge representation is very important for Rough-Set data processing. Data are stored in a decision table. The columns represent attributes and the rows represent objects whereas every cell contains attribute value for corresponding objects and attributes. Decision tables are also called information systems. A decision table (*DT*) is the quadruple $T = (U, A, C, D)$, where U is a nonempty finite set of objects called the universe, A is a nonempty finite set of primitive attributes, and $C, D \subseteq A$ are two subsets of attributes that are called the condition and decision attributes.

A unique feature of the RST method is its generation of rules that played an important role in predicting the output. Rosetta listed the rules and provides some statistics for the rules which are support, accuracy, coverage, stability and length. Below is the definition of the rule statistics (Bose, 2006)

i) The rule LHS support is defined as the number of records in the training data that fully exhibit property described by the IF condition.

ii) The rule RHS support is defined as the number of records in the training data that fully exhibit the property described by the THEN condition.

iii) The rule RHS accuracy is defined as the number of RHS support divided by the number of LHS support.

iv) The rule LHS coverage is the fraction of the records that satisfied the IF conditions of the rule. It is obtained by dividing the support of the rule by the total number of records in the training sample.

v) The rule RHS coverage is the fraction of the training records that satisfied the THEN conditions. It is obtained by dividing the support of the rule by the number of records in the training that satisfied the THEN condition.

vi) The rule length is defined as the number of conditional elements in the IF part.

Australian Journal of Basic and Applied Sciences, 8(4) Special 2014, Pages: 114-122

3. Rough Set Methodology For Classifying Cardiac Dataset:

When a classifier is presented with a new case, the rule set is scrutinized to find pertinent rule that is the rules that the predecessors match the case. If no rule is found, the most frequent outcome in the training data is chosen. If more than one rules match, these may in turn indicate more than one possible outcome. A voting process is executed across the matched rules to resolve the conflicts and to rank the predicted outcomes. This study employed cardiac dataset from National Heart Institute (NHI), Malaysia database collected from 1997 till 2003. The cardiac dataset consists of 6,892 records, consequences from three forms that are used by NHI;

i) Hospital Notes: Initial Consult Cardio Thoracic Surgery
ii) Hospital Notes: Operative Cardio Thoracic Surgery
iii) Hospital Notes: Discharge Cardio Thoracic Surgery

All the above sheets record the data for each form of NHI Hospital Notes. The attributes in this data describes a medical details of cardiac patient which will determine the complication in the hospital that cardiac patient will have after operation. 14 attributes describe the condition of cardiac patient and one attribute describe the decision of 'complication'; the early warning signal for the doctor to prepare with the subsequent procedure after the patient undergoes cardiac operation. The real cardiac set collection is mounting continuously as more and more cases are analyzed and recorded. Table I are the attributes description for NHI cardiac dataset.

Table I: Description Of Attributes For Cardiac Data

No	Attribute	NHI sheet	Description
1	age year	Initial Consult	Patient's age in year unit.
2	morbid obesity	Initial Consult	If patient's weight more than 1.5 times of BMI, it consider as obesity.
3	diabetes	Initial Consult	Patient that have diabetes or patient using the anti-diabetic medicine.
4	hypertension	Initial Consult	Patient with hypertension.
5	operative	Initial Consult	How many operation that patient have before.
6	dialysis	Initial Consult	Patient that have hemodialysis or peritoneal dialysis before having the operation.
7	smoking history	Initial Consult	Patient is smoker or not.
8	family CAD	Initial Consult	Family history of cardiac disease.
9	hypercholestrol	Initial Consult	Patient that have hypercholestrol.
10	renal failure	Initial Consult	Patient that have renal disorder.
11	cerebroaccident	Initial Consult	Patient that have brain accident.
12	infectious endocarditis	Initial Consult	Patient that have infectious endocarditis.
13	peripheral vascular	Initial Consult	Patient that have peripheral vascular
14	immunosuppresive	Initial Consult	Patient that have immunosuppression.
15	complication	Discharge	Decision attribute which is the complication that patient will have in hospital.

4. Generation of Significant Reducts and Rules:

The significant reduct are based on core attributes and minimal cardinality. The core attributes is the set of attributes which is common to all reducts. The core is the set of attributes which is possessed by every legitimate reduct, and therefore consists of attributes which cannot be removed from the information system without causing collapse of the equivalence class structure. The intersection of all reducts is called the core reduct; the elements of attributes that cannot be eliminated as well as the s*et al*l indispensable attributes. The core is defined as;

$$\text{Core}(C) = \cap \text{Red} \quad (1)$$

In rough set theory, all of indispensable attributes should be restricted in an optimal attribute subset. Core is the s*et al*l indispensable attributes. The process of searching indispensable attributes is that of finding the CORE.

In rough set attribute reduction, a reduct with minimal cardinality is searched for. An effort is made to locate a single element of the minimal reduct set Red*d*min □ R*ed* The reduct with minimal cardinality is the reduct with minimal length.

$$\text{Red}\min = \{R \square \square \text{Red} \mid \vee A' \square \square \text{Red}, |R| \leq |R'|\} \quad (2)$$

The significant rules are based on rules support, rules length and rule important measure (RIM). Given a description contains a conditional part α and the decision part β, denoting a decision rule $\alpha \rightarrow \beta$. The support of the pattern α is a number of objects in the information system A has the property described by α.

$$\text{Sup}\ port(\alpha) = ||\alpha|| \quad (3)$$

The support of β is the number of object in the information system *A* that have the decision described by β.

Australian Journal of Basic and Applied Sciences, 8(4) Special 2014, Pages: 114-122

$$\text{Sup } port(\beta) = ||\beta|| \quad (4)$$

The support for the decision rule $\alpha \rightarrow \beta$ is the probability of that an object covered by the description is belongs to the class.

$$\text{Sup } port(\alpha \rightarrow \beta) = \text{Sup } port(\alpha \,.\, \beta) \quad (5)$$

For the accuracy measurement, the quantity accuracy ($\alpha \rightarrow \beta$) gives a measure of how trustworthy the rule is in the condition β. It is the probability that an arbitrary object covered by the description belongs to the class. It is identical to the value of rough membership function applied to an object x that match α. Thus accuracy measures the degree of membership of x in X using attribute B.

$$\text{Accuracy } (\alpha \rightarrow \beta) = \frac{\text{Sup } port(\alpha \,.\, \beta)}{\text{Sup } port(\alpha)} \quad (6)$$

Coverage measurement measures the behavior of pattern α in describing the decision class defined through β. It is a probability that an arbitrary object, belonging to the class C, and is covered by the description D.

$$\text{Coverage } (\alpha \rightarrow \beta) = \frac{\text{Sup } port(\alpha \,.\, \beta)}{\text{Sup } port(\beta)} \quad (7)$$

The rules are said to be completed if any object belonging to the class is covered by the description coverage is 1, while deterministic rules are rules with the accuracy is 1. The correct rules are rules with both coverage and accuracy is 1.

Rules generated from reduct are representative rules extracted from the data set. Since a reduct is not unique, rule sets generated from different reducts contain different sets of rules. However, more important rules will appear in most of the rule sets. Less important rules will appear less frequently than those more important ones. Some rules are generated more frequently than the others among the total rule sets. Such rules are considered as more important rules. The Rule Importance Measure (RIM) is computed according to the frequency of an association rule among the rule sets. The Rule Importance Measure (RIM) is defined as follows,

$$\text{Rule Importance Measure} = \frac{\text{Frequency of Appeared Rules from Reduct Set}}{\text{Number of Reducts Set}}$$

In this study, the following tasks have been done by using NHI cardiac dataset:

i) Missing values of the dataset have been removed by incorporating the incomplete process.
ii) Dataset are split into 70% of training and 30% of testing records
iii) Training dataset has gone through the Equal Frequency Binning Discretization.
iv) The discretization data generates the reducts.
v) The rules are produced for classification process.
vi) An analysis and evaluation of:
a) The generated reduct by
i. Choosing minimal cardinality.
ii. Core attributes in the generated reduct are analyzed.
b) A new decision table is constructed based on the attributes consist of reducts with minimal cardinality and core attributes
vii) New decision table have the same process as no i) to v)
viii) Second phase of analysis and evaluation:
a) Generated Rules are analyzed accordingly;
i) Rules with highest support values are chosen
ii) Rules with less length are preferred
iii) Rules with highest percentage of Rule
Importance Measure (RIM) are favored
ix) Classification process is done as in step (vi)
x) Results of classification are compared in terms of classification accuracy of original decision table and new decision table.
xi) Experimental results are analyzed and discussed.

Australian Journal of Basic and Applied Sciences, 8(4) Special 2014, Pages: 114-122

5. Experimental Result and Analysis:

Cardiac dataset consists of 6892 records with no missing values. The data are divided into two parts; training and testing group. The training group is split into 70% which equal to 4824 records, while the testing group is accounted for 30% which equal to 2068 records. The testing group consists of 2068 records, and the classification is implemented using standard voting classifier. The derived rules from the training phase are used to test the effectiveness of the unseen data.

Training data is discretized using EFB to obtain an equal number of objects into each interval. The interval is determined by the number of bin, *n*-1. Several numbers of bins are tested in this study to obtain high classification rates (Refer to Table 3). From table II , it shows by using few bins, the classification rates are improved. Number of bins that equal to 3 yields 74% classification accuracy, while large number of bins, i.e., *n*=20 gives 50% classification accuracy. Therefore, this study adopts number of bin equals to 3 for reduct generation and rules derivation.

Table II: Comparison Of Classification With Different Numbers Of Bins.

Number of Bin	Classification Accuracy (%)
Bin 3	74%
Bin 5	60%
Bin 10	53%
Bin 20	50%

Genetic Algorithm is used for reduct generation as it provides more exhaustive search of the search space. Reduct with object related is used, which produce a set of decision rules or general pattern through minimal attributes subset that discern on a per object basis. The reduct with object related have capability in generating reduct based on discernibility function of each object.

Based on the generated reducts with length 1 and 2, the core attributes are {morbid obesity, dialysis, smoking history, renal failure, cerebroaccident and infectious endocarditis}. These attributes are important attributes to obtain better classification in testing phase. The reduct with minimal cardinality also contribute to the connotation reduct in generating the significant rule. It will consider the reduct with minimal cardinality of minimal length. In this experiment, the reduct with minimal cardinality is {infectious endocarditis} with length of 1. Based on these core attributes and attributes with minimal cardinality, new decision table are mapped. Subsequently, the rules generated from this new table are analyzed for better classification compared to original NHI cardiac dataset without prior analysis on reduct and rules generated.

Table 4 is a new description of new decision table for NHI cardiac data. Original NHI cardiac data have 15 attributes including the decision attributes. Nevertheless, the new decision table based on the analyzed reducts reveals 6 attributes. These new rules derivation are analyzed based on the rough set benchmark and measurement for better classification than original decision table of NHI cardiac data.

To uncover the most significant rules, these rules are sorted according to their support value. The highest support value is resulted as the most significant rules. All rules are generated with statistics rule. Based on the sorted of highest rule support values, the most significant rule is {dialysis(0) AND cerebroaccident(0) AND infectious endocarditis(0) => complication(0) OR complication(1)} with the outcome of no complication (output = 0) and with complication (output=1). This is supported by 4758 for LHS support and 4228 and 530 for RHS support value. The RHS support values have two different values, depending on the numbers of records in the training dataset described by the THEN condition; complication (0) or complication (1).

Subsequently, the impact of rules length on testing accuracy are evaluated based on rules set from new NHI Cardiac Data Decision Table (refer to table III). Consequently, the same rules are divided into two groups; rules of length ≤ 3 and rules of length > 3. It seems that the rules with length ≤ 3 contribute better classification compared to the rules with length > 3.

Table IV: Effect Of Rules Length On Testing Dataset

Rules of Length ≤ 3	Rules of Length > 3
89.2%	86.8%

This section demonstrates the analysis of the generated rule. The rule statistic involve in this analysis are rule support, rule coverage and rule accuracy. Rules computations are done for rule 1 through rule 10. Table IV illustrates the rule computations based on rule derivation.

Based on Table IV, Rule {infectious endocarditis(1) => complication(0)} is taken as illustration of rules computation, and this is denoted as $(\alpha \rightarrow \beta)$.

Table V: Rule Computations

Rules	LHS Support	RHS Support	RHS Accuracy	LHS Coverage	RHS Coverage
infectious endocarditis(1) => complication(0)	10	10	10/10 = 1.0	10/4823 = 0.002073	10/4286 = 0.002333
dialysis(1) AND renal failure(1) => complication(0) OR complication(1)	7	4,3	4/7 = 0.571429, 3/7 = 0.428571	7/4823 = 0.001451	4/4287 = 0.000933, 3/4287 = 0.005576
dialysis(1) AND renal failure(0) => complication(0)	3	3	3/3 = 1.0	3/4823 = 0.000622	3/428 = 0.0007
renal failure(1) AND cerebroaccident(1) => complication(0)	2	2	2/2 = 1.0	2/4823 = 0.000415	2/4282 = 0.000467
morbid obesity(1) AND renal failure(1) => complication(0)	1	1	1/1 = 1.0	1/4823 = 0.000207	1/4291 = 0.000233

α is LHS condition and β is RHS condition. The computation procedure of $(\alpha \rightarrow \beta)$ for support, accuracy and coverage are shown below.

i) Support$(\alpha \rightarrow \beta)$ = Support$(\alpha . \beta)$
= Support (10.10)
= 20

ii) Accuracy $(\alpha \rightarrow \beta)$ = $\frac{\text{Support}(\alpha . \beta)}{\text{Support}(\alpha)}$
= $\frac{\text{Support (10.10)}}{\text{Support (10)}}$
= 10

iii) Coverage $(\alpha \rightarrow \beta)$ = $\frac{\text{Support}(\alpha . \beta)}{\text{Support}(\beta)}$
= $\frac{\text{Support (10.10)}}{\text{Support (10)}}$
= 10

Rule Importance Measure (RIM) is used to evaluate the importance of association rules. The analysis of RIM is demonstrated to determine the importance of the attributes. The number of reduct set generated from new decision table is 14 and the attributes of renal failure have the highest RIM percentage which is 64.2% (refer to Table V). Consequently, all the rules in renal failure attributes are chosen for classification process, thus contributing to better classification accuracy.

Table VI: Rim Using Rules From New Decision Table

No.	Attributes	RIM (%)
1	morbid obesity	4/14 = 28.6%
2	dialysis	4/14 = 28.6%
3	smoking history	2/14 = 14.2%
4	renal failure	9/14 = 64.2%
5	cerebroaccident	5/14 = 35.7%
6	infectious endocarditis	4/14 = 28.6%

The significant rules are determined based on;

i) the rule with highest number of support value,
ii) the rule with less length, and
iii) the rule with highest number of RIM percentage.

These significant rules are conceded for classification process to improve the classification (refer Table VI for complete generated significant rules).

From previous analysis, it reveals that better classification has been achieved. Generally, physicians do not need to use all the attributes and rules to diagnosis the patients since these will incur long processing time, and no guarantee of better performance. Hence, the core attributes and the significant rule are preferred for quick decision making in determining better result for classification. Table VII illustrates the result of classification performance of original NHI cardiac data and the new decision table of NHI cardiac data. Figure 1 illustrates the classification accuracy for original decision table and new decision table of NHI cardiac dataset.

Table VII: Significant Rules Of Cardiac Data

Rules	LHS Support	RHS Support	LHS Length	RHS Length
dialysis(0) AND cerebroaccident(0) AND infectious endocarditis(0) => complication(0) OR complication(1)	4758	4228, 530	3	2
morbid obesity(0) AND dialysis(0) AND renal failure(0) AND infectious endocarditis(0) => complication(0) OR complication(1)	4683	4166, 517	4	2
morbid obesity(0) AND renal failure(0) AND cerebroaccident(1) AND infectious endocarditis(0) => complication(0) OR complication(1)	44	39, 5	4	2
morbid obesity(0) AND smoking history(0) AND renal failure(0) AND cerebroaccident(1) => complication(0) OR complication(1)	27	25, 2	4	2
renal failure(1) AND cerebroaccident(0) => complication(0) OR complication(1)	69	55, 14	2	2
renal failure(1) AND cerebroaccident(1) => complication(0)	2	2	2	1
infectious endocarditis(1) => complication(0)	10	10	1	1
morbid obesity(1) AND renal failure(1) => complication(0)	1	1	2	1
smoking history(1) AND renal failure(1) => complication(0) OR complication(1)	30	23, 7	2	2
dialysis(1) AND renal failure(1) => complication(0) OR complication(1)	7	4, 3	2	2
dialysis(1) AND renal failure(0) => complication(0)	3	3	2	1

Table VIII: Classification Performance Of Both Nhi Cardiac Data

Decision Table	Rule set	Overall Accuracy
Original decision table	All rules	74.8%
New decision table	Selected rules	89.6%

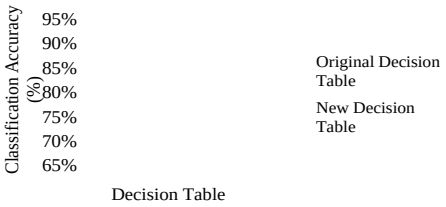

Fig. 1: Accuracy for original decision table and new decision table for NHI cardiac dataset

A new generation of decision table for NHI cardiac dataset gives a significant impact to the classification rates. Reduct and rules analysis have yielded significant attributes and rules, thus proven to have better classification accuracy compared to the results which employ all attributes, reduct and generated rules. This improvement is also supported by the number of bins in EFB discretization. The less number of bins tend to give better classification accuracy compared to large number of bin.

6. Comparison With Other Medical Dataset:

We conclude that, the accuracy of different medical datasets depends on the behavior and characteristics of the data (refer to table VIII). Since in this study, we are dealing with actual cardiac datasets from NHI which might have some noise, the highest accuracy is obtained selected rules. Hence, these rules can be an important message for the physicians to give early information to the patients accordingly.

Table IX: Comparison Of Classification Accuracy Obtained By Using Different Classifiers For Medical Datasets

No	Methods	Dataset	Accuracy (%)	Reference
1	Rough Set + SVM	Breast Cancer	99.41	(Chen *et al*., 2011)
2	Fuzzy Sets and Rough Sets + Statistical	Breast Cancer	98.46	(Hassanien, 2007)
3	Rough Set + PSO	Brain Glioma	86	(Wanga *et al*., 2006)
4	Classical Rough Set + SVM-RBF	Colon	87.10	(Wang *et al*., 2011)
5	Classical Rough Set + SVM-RBF	Leukimia	97.22	(Wang *et al*., 2011)
6	Wavelet Packet Transforms + Neighborhood Rough Set + SVM	Colon	95.51	(Zhang *et al*., 2010)
7	Wavelet Packet Transforms + Neighborhood Rough Set + SVM	Leukimia	90.46	(Zhang *et al*., 2010)
8	Fuzzy + Rough Set + C5.0	Colon	91.9	(Xu *et al*., 2009)
9	Fuzzy + Rough Set + C5.0	Leukimia	98.6	(Xu *et al*., 2009)

Conclusion:

An attempt has been made in this study to explore the significant of reduct and rules that contributing to better classification performance. This study has presented a detail methodology to devise a framework of Rough Set Cardiac Data. The process involves a set of procedure principally for reduct generation, rules derivation and classification. Owing to Rosetta flexibility, rough set technique can be applied to the cardiac medical dataset. Several analyses have been achieved to find the significant reduct and rules for better classification. Consequently, the significant attributes are analyzed based on the minimal cardinality and the core attributes of the generated reduct. A new decision table of NHI cardiac dataset has been constructed. The rules generated based on this new decision table are analyzed based on highest support value, rules with less length and Rule Importance Measure (RIM). Nevertheless, the rules with less length cannot establish the significant rules, since the degree of significant depends on high support value for the rules that are being analyzed. As a result, in this study, the influences of using core attributes with minimal cardinality in the course of the generated reduct and significant rules have been examined. An empirical study has been conducted for searching optimal classification. A rough set framework for cardiac medical dataset is illustrated mutually with an analysis of reduct and derived rules, with entrenchment of their implicit properties for better classification outcomes.

From the experiments and the acquired results, it depicts that Rough set is a remarkable soft computing technique for handling medical data with the existence of missing values. The reduct with core attributes and minimal cardinality assists better classification enhancement. The rules with less length are not efficient as a rule significant measurement. The rules derivation with highest support value, less length and high value or Rule Importance Measure (RIM) are proven to be significant rules in contributing to better classification. The significant of reduct and rules are required to produce better classification result.

ACKNOWLEDGEMENT

Authors would like to thank to *Faculty of Computer, Media and Technology Management (FKMPT),* TATI University College and *Soft Computing Research Group (SCRG),* Universiti Teknologi Malaysia for the support in making this study a success.

REFERENCES

Andrew, W., 1999. “Statistical Pattern Recognition” Oxford Univeristy Press Inc., Oxford

Bose, I., 2006. “Deciding the Financial Health of Dot-Coms Using Rough Sets. *Information Management.* 43(7): 835-846

Hassanien, A.E., 2007. “Fuzzy rough sets hybrid scheme for breast cancer detection” Quantitative Methods and Information Systems Department, College of Business Administration, Kuwait

Chen, H.L., B. Yang, J. Liu and D.Y. Liu, 2011. “A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis” College of Computer Science and Technology, Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, China.

Narendra, P.M. and K. Fukunaga, 1977. ”A Branch and Bound Algorithm for Feature Subset Selection” IEEE Transactions on Computers., 26(9): 917-922.

Øhrn, A., 1999. “Discernibility and Rough Sets in Medicine: Tools and Applications” PhD Thesis, Department of Computer and Information Science, Norwegian University of Science and Technology, Trondvwhefwddfvr fvheim, Norway.

Pawlak, Z., 1982. “Rough Sets. International Journal of Information and Computer Sciences” 11: 341-356.

Pawlak, Z., 1991. “Rough Sets In Theoretical Aspects of Reasoning about Data” Kluwer Academic Publishers.

Pawlak Z., 1998. “Rough Set Theory and its Applications to Data Analysis” Cybernetics and Systems, 29: 7 661-688.

Quinlan, J.R., 1986. “Induction of Decision Trees - Machine Learning 1: 81-106” Kluwer Academic Publishers.

Su, C. and C. Yang, 2008. ‘’Feature selection for the SVM: An application to hypertension diagnosis” Expert Syst. Appl., 34(1): 754-763

Wang, S.L., H.W. Chen, F.R. Li, Zhang and D.X., 2011. “Gene selection with rough sets for the molecular diagnosing of tumor based on support vector machines” International Computer Symposium, Taiwan., pp: 1368-1373.

Wang, X., J. Yang, X. Teng, W. Xia and R. Jensen, 2007. “Feature selection based on rough sets and particle swarm optimization” Pattern Recogn. Lett., 28(4): 459-471g.c

Wanga, X.Y., J. Yanga, R. Jensen and X.J. Liua, 2006. "Rough set feature selection and rule induction for prediction of malignancy degree in brain glioma" Institute of Image Processing and Pattern Recognition, Jiao Tong University, China, Department of Computer Science, The University of Wales, UK.

Xu, F.F., D.Q. Miao and L. Wei, 2009. "Fuzzy-rough attribute reduction via mutual information with an application to cancer classification" Department of Computer Science and Technology, Key Laboratory of Embedded System and Service Computing, Ministry of Education of China, Tongji University, China.

Ye C.-Z., J. Yang, D.-Y. Geng, Y. Zhou and N.-Y. Chen, 2002. "Fuzzy rules to predict degree of malignancy in brain glioma" Med. Biol. Eng. Comput. 40.

Zhang, S.W., D.S. Huang, S.L. Wang and 2010. "A method of tumor classification based on wavelet packet transforms and neighborhood rough set" Institute of Intelligent Machines, Chinese Academy of Sciences, China

Zurada and M. Jacek, 1992. "Introduction to Artificial Neural Systems" West Publishing Company.