CHOOSING THE BEST KNOWLEDGE ACQUISITION FOR DENGUE DATASET BASED ON ROUGH SET APPROACH

[1]Mohamad Farhan Mohamad Mohsin*, [2]Mohd Helmy Abd Wahab†, [3]Azuraliza Abu Bakar†
[1]College of Arts & Sciences,Universiti Utara Malaysia, 06010 UUM Sintok, Kedah, Malaysia
[2] Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia
[3] Faculty of Science and Information Technology, University Kebangsaan Malaysia, 43600 UKM Bangi, Malaysia.
[1]farhan@uum.edu.my, [2]helmy@uthm.edu.my, [3]aab@ftsm.ukm.my

ABSTRACT

Knowledge based system need quality knowledge to provide an accurate decision. Due to irrelevant and incorrect information of expert during knowledge acquisition, the information extraction from data mining approach can reduce the biases and misconceptions of multiple experts. Moreover, the real world data which is huge and incomplete has made it complicated to extract only the quality knowledge. Hence, this paper is to investigate the capability of rough set (RS) in generating quality knowledge for dengue dataset problem. Four RS reduct algorithms i.e Genetic Algorithm (GA), Johnson Reducer (JR), Exhaustive Calculation (EC) and Dynamic Reduct (DR) are compared. The comparisons are made based on the accuracy, rules quantity, and rules length. The dengue dataset contains 182 is obtained from Sungai Petani Hospital, Kedah, Malaysia. The result shows that the performances of all reduct algorithms have comparable accuracy and in several algorithms, they obviously generates more and longer rule. GA is the most accurate classifier with shorter rule. Similarly, JA produces comparable accuracy but distinctly generate lower number and shorter rule. EC and DR are capable to produce good accuracy but use more and longer rules to obtain the good accuracy. In conclusion, JA is chosen as the best method for knowledge acquisition for dengue dataset.

Keywords : Rough set, reduct function, knowledge quality, dengue

INTRODUCTION

Knowledge based system (KBS) need quality knowledge in order to provide an accurate decision. The acquisition of the valuable knowledge is practically obtained from experts of specific domain. This approach has been practice in many intelligent systems applications and it becomes the main constraint in extracting knowledge due to irrelevant and incorrect information of expert [1]. Presently, this process can be automated by data mining technique which can produces information similar as an expert [2]. The extracted information from data mining approach can reduce the biases and misconceptions of multiple experts [3]. Moreover, the rapid advance of data storage makes database contain so much data which offers a high probability to extract more interesting knowledge. Based on the assumption “people can lie but data not”, the dataset can be a very useful for knowledge acquisition using data mining approach [4].

“Rule” generated from any classification system is considered as knowledge which can be used in KBS [5]. In the real world application, the number of attribute of a dataset could be very large. Following that, the large dataset may contain thousands of relationship and it will likely provide more knowledge since the interrelationship between data will give more description [6]. Furthermore, it is also have the possibility to have most number of rules that contain unnecessary rule or redundancies in the model. Because of that, the search for appropriate data mining approach which can provide quality knowledge is important.

Theoretically, a good set of knowledge should provide good accuracy when dealing with new cases. Besides accuracy, a good rule set must also has a minimum number of rules and each rule should be short as possible. It is often that a rule set contains smaller quantity of rules but they usually have more conditions. An ideal model should be able to produces fewer, shorter rule and classify new data with good accuracy. The quality and compact knowledge will contribute manager with a good decision model. According to [7], the best model should be simple, easy to understand, robust, and can be adapt with new information The knowledge quality is an important criteria in order to achieve the purpose of KBS; to create a reliable method of making predictions and recommendation.

Rough Set (RS) is one of data mining task which is capable to provide knowledge for KBS. It is generally used in classification problem and has high capability to generate knowledge via rule from imprecise information. Since the data in the real world problem always incomplete, RS is almost the suitable method to search for a set of quality knowledge. Hence, this paper is aimed to investigate the most suitable approach in rough set that capable to generate quality knowledge for dengue data set. In this study, four reduct algorithms of RS i.e Genetic Algorithm[8], Johnson Reducer [9], Exhaustive Calculation, and Dynamic Reduct [10] which control the quality of rule will be compared in order to determine the best knowledge acquisition for dengue data set. The purpose of this comparison is to examine the classifier accuracy and knowledge quality in term of rule quantity, and rule length. The finding of this study is helpful enough for the construction of dengue knowledge-based applications in the future.

This paper is organized as follows. Section 2 will outline the theory of RS. The model development is discussed in section 3. The experiment and result will be presented in section 4 and final section will conclude this work.

ROUGH SET THEORY

In this section, the theory of *RS* will be discussed. *RS* theory was developed by Zidslaw Pawlak in the early 1984s, provides a mathematical approach to deal with uncertainty and vagueness [11]. The main goal of the approach is to derive rules from a set of data which is represented in a decision system. It is donated as $D_s= \{U, D_c, D_d\}$ where U is a finite set of objects called *universe*, D_c is a set of condition attributes and D_c is decision attribute. *RS* theory is based on the establishment of equivalence classes in training data set and the formation is almost indiscernible. *RS* deals with imprecise information in information system (*IS*) using the concept of set approximation which each vague object are described in lower and upper approximation space. Consider an $IS= \{U, A\}$ and $X|\overline{U}$ be a selected of attributes. B-lower approximation is defined as $\underline{B}x=\{X \in U:[x]_B \subseteq X\}$ while B-upper approximation is $Bx=\{X \in U:[x]_B \cap X \neq 0\}$. The set $\underline{B}x$ (or Bx) consists of objects which are surely belonging to X which respect to the knowledge provide by B. The set $BN_B(X) = \underline{B}x\text{-}Bx$ is called B-boundary of X which consist of those objects that not surely belong to X.

Rough set modeling assumes the existences of several attribute in a data set are more important compares to others. This is done by the reduct function which will determine only important attribute to represent the whole problem. Reduct calculation determines the number rule, rule length, and accuracy of the classifier. There are several algorithms to perform reduct such as Johnson Reducer, Genetic Algorithm, Dynamic Reduct, Holte1R, and SIP/DRIP. Each method use different method to control the generation of rule from dataset. Generally, the rough set approach consists of several steps leading towards the final goal of generating rule from information system as given below [12]:

- Forming decision system by mapping information from the data source
- Completion of data
- Descritization of data
- Reducts computation of data
- Obtain rule from the reduct
- Classification of new unseen data

MODEL DEVELOPMENT

Dengue data profile which ranges from the year 2004-2005 was selected as data experiment. It was manually collected from a medical report of suspected dengue patients from Sungai Petani Hospital, Kedah, Malaysia. The original dataset had 182 cases, each described by 37 attributes including target class (2 numerical, 4 continuous, 31 nominal). The target class holds two classified groups that are positive dengue and negative dengue. In preliminary observation the raw dataset contains missing values and outliers.

Model development consists of two parts i.e. data preprocessing and rule extraction. In data preprocessing, dengue dataset was pre-processed where all unknown numeric attributes were replaced with mean value while max value for character attributes. After preprocessed, the origin numbers of attributes were reduced to 17 attributes and a target class. After that, the data were discretized using boolean reasoning technique [13]. The data were then split into training and testing using *n*-fold cross-validation technique where 9 folds of data are prepared based on ratio of training and testing; 10:90, 20:80, 30:70 40:60, 50:50, 60:40, 70:30, 80:20, and 90:10. In rule extraction phase, rough set data analysis tool namely Rosetta v3.2 was selected to generate rule and perform classification. Four reduct algorithms embedded in Rosetta was utilized to extract rule. During experiments, the accuracy of the classification, the number of rule, and the length of rule were recorded. Figure 1 depicts the modeling process in RS.

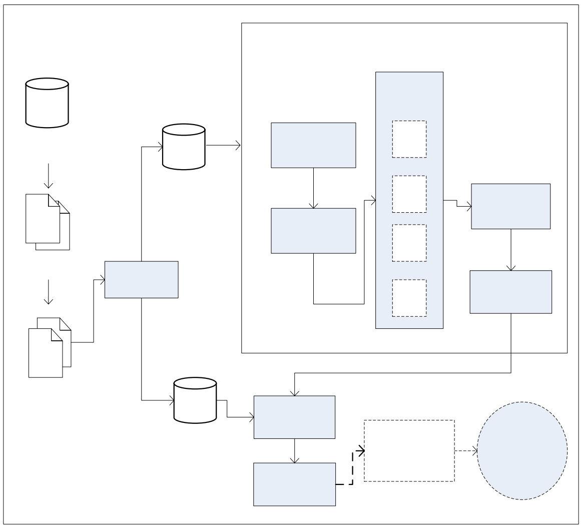

Figure 1. Modeling Processes in RS.

EXPERIMENT AND RESULT

This section will discuss the experimental results of RS on dengue dataset. The evaluation is based on the performance of four reduct functions in RS which is focused on accuracy, number rule, and rule length. The goals and notations of the experiments are formalized. The rough set model is given as ***RS***. The training and testing data are given as ***tr*** and ***ts***. The tested reduct functions namely Genetic Algorithm, Johnson Reducer, Exhaustive Calculation and Dynamic Reduct are given as ***GA***, ***JR***, ***EC***, and ***DR***. The accuracy, number of rule, and length of rule are given as ***acc***, ***nr***, ***rl***. The *acc* of the classification shows how good the obtained rule can classify new cases. Beside that, the *nr* indicates the amount of knowledge needed to provide good accuracy while the *rl* is number of condition attributes in the rule. Table 1(a-c) depict the experimental results of reduct function on dengue data set. The ***b_acc***, ***b_nr***, and ***b_rl*** in table 1(c) are the highest *acc*, lowest *nr*, and shortest *rl* among GA, JR, EC, and DR in 9 models.

Table 1(a). Experimental results of GA, JR, EC, and DR for dengue dataset (cont.)

tn:ts	*GA*			*JR*		
	acc	*nr*	*rl*	*acc*	*Nr*	*rl*
90:10	58.3	9157	7	66.7	103	4
80:20	66.7	8328	5	69.4	94	4
70:30	69.1	7202	6	65.5	82	5
60:40	69.8	6099	5	67.1	72	4
50:50	66.0	5664	5	61.5	66	4
40:60	70.0	3993	5	67.1	55	4
3070	63.8	2683	4	62.9	39	3
20:80	56.8	1482	4	63.0	25	3
10:90	72.2	873	4	45.1	14	2

Table 1(b). Experimental results of GA, JR, EC, and DR for dengue dataset (cont.)

tn:ts	*EC*			*DR*		
	acc	*nr*	*rl*	*acc*	*nr*	*rl*
90:10	66.7	29844	8	66.7	109335	8
80:20	63.9	25723	8	62.8	90335	8
70:30	70.9	21175	8	70.9	73988	8
60:40	67.1	17451	8	71.2	57795	8
50:50	60.4	15761	7	57.1	48604	7
40:60	64.2	10111	8	66.9	31041	8
3070	63.7	6445	7	63.7	10041	7
20:80	54.8	3017	7	65.7	7602	7
10:90	57.9	1177	5	61.6	2378	5

Table 1(c). Experimental results of GA, JR, EC, and DR for dengue dataset (cont.)

tn:ts	*b_acc*	*b_nr*	*b_rl*
90:10	GA	JR	JR
80:20	JR	JR	JR
70:30	DR	JR	JR
60:40	DR	JR	JR
50:50	GA	JR	JR
40:60	GA	JR	JR
3070	GA	JR	JR
20:80	DR	JR	JR
10:90	DR	JR	JR

As shown in *acc* column in table 1 (1-b), the *acc* of the classifications are averagely low. This occurs because of the data problem itself when there are many conflicts between positive and negative dengue symptoms inside the data. Generally, the results indicate that the *acc* of all reduct algorithms are comparatively scored at similar ranges. However, the *nr* and *rl* are seemed significantly different. From table 1(c), JR is found generated lower *nr* and shorter *rl* for all experiment while in *acc*, GA, JR, and EC comparatively lead the highest *acc* in certain model. The whole experiments are summarized in table 2. Table 2 represents the average acc (***avg_acc***), highest accuracy (***max_acc***), average nr (***avg_nr***), the lowest nr (***best_nr***), and shortest rl (***best_rl***).

Table 2. Summarization of the experiment

	avg_ acc	***max_ acc***	***avg_ nr***	***best_ nr***	***best_ rl***
GA	65.9	72.2 (90:10)	5053	873(10:90)	4
JR	63.1	67.1 (3070)	61	14 (10:90)	2
EC	63.3	70.9 (70:30)	14523	1177 (10:90)	5

DR	65.2	71.2 (60:40)	47902	2378 (10:90)	5

The *avg_acc* in table 2 determines the *acc* distribution among *ts* data set. In *avg_acc* column, GA turns out to be the most accurate classifier when obtains highest score in all models. Moreover, the 72.2% of GA *max_acc* is achieved in 10:90 model which indicates the ability of GA to classify more unseen cases (90% of the total records) using limited number of rule (*nr*=1177) compared to other method. Others methods- DR, EC, and JR has comparable *avg_acc* which roughly less 2.8% accurate than GA. Figure 2 depicts the *avg_acc* of GA, JR, EC, and DR for dengue dataset.

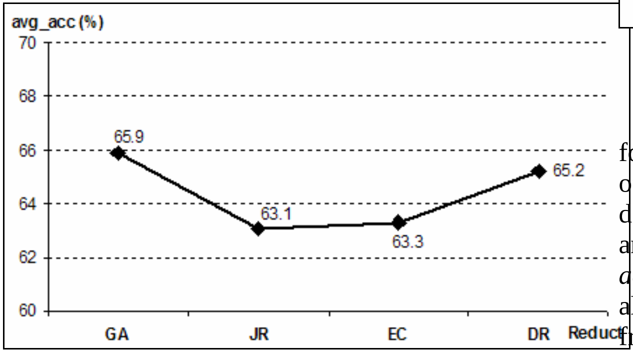

Figure 2. Avg_acc of GA, JR, EC, and DR for dengue dataset.

.

In term of *nr*, JR outperform as lowest *nr* generator. Out of 9 models, JR generates only 61 rules in average. This is followed by GA which generates more rules compared to JR. However, for EC and DR methods, they obviously generates huge *nr* with significant different with other methods. In average, EC and DR, generate 14523 rules and 47902 rules which are 100% more than JR and GA. From the observation, the *nr* is increased when more *tn* data presented to them. Similar to *nr*, the number of condition of the rule or *rl* is also increased when more *tn* involve during mining. From table 1 (a) and table 1(b), GA and JR comparatively generates shorter rule while *rl* for DR and EC are much longer. If the *rl* is lower, it will increase the robustness of the knowledge to handle more unseen cases.

Typically, the *nr* and *rl* will leads the classifier to obtain good *acc* result. If there are two models with equal *acc*, the model with lower *nr* and *rl* will be chosen as better system since it can performs the same task using limited knowledge. Figure 3 shows the relationship between *avg_acc* and and *avg_nr* of GA, JR, EC, and DR for dengue dataset. From the figure, JR only needs 61 rules to provide 63.1% *acc*. For EC and DR, they achieve higher *acc* than JR nevertheless both of them require more rule to provide the comparative result.

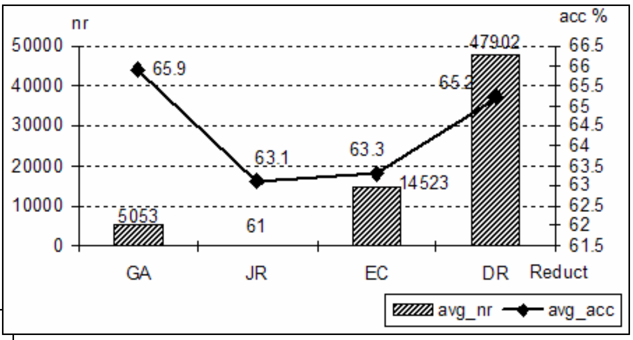

Figure 2. Avg_acc of GA, JR, EC, and DR for dengue dataset.

From the experiments, the best reduction method for dengue dataset can be determined. Frequently, *acc* of the classifier will be chosen as a main criterion to decide the best model. In this study, several criterions are used as the best model characteristic that are higher *acc*, fewer *nr*, shorter *rl*, and the quantity of data allocation for *tn* and *tst* If the model generates fewer *nr* from limited number of *tn* dataset but can classify more *ts* dataset with high *acc*, it might be the best model. Thus, the best models are classified into three groups that are the best model in term of acc (***b$_M$_acc***), the best model in term of *nr* (***b$_M$_nr***), and the best model in term of *rl* (***b$_M$_rl***). The best model of GA, JR, EC, and DR are depicted in table 3. The right column of table 3 shows the *b$_M$_acc*, *b$_M$_nr*, and *b$_M$_rl* of the best selected model. The general formulations to choose the best model to obtain the best model of reduct function are listed as follows:

- $\forall R_x$, *if acc* $(R_x \rangle R_{x+1}) \rightarrow R_x \rangle_{\text{better}} R_{x+1}$
- $\forall R_x$, *if nr* $(R_x \langle R_{x+1}) \rightarrow R_x \rangle_{\text{better}} R_{x+1}$
- $\forall R_x$, *if rl* $(R_x \langle R_{x+1}) \rightarrow R_x \rangle_{\text{better}} R_{x+1}$

where *R* is a reduct and *x = {GA, JR, EC, and DR}*

Table 3: The b$_M$_acc, *b$_M$_nr*, and *b$_M$_rl* of GA, JR, EC, and DR for dengue dataset

	tn:ts	*acc*	*nr*	*rl*	***b$_M$*_acc**	***b$_M$*_nr**	***b$_M$* _rl**
GA	10:90	72.2	873	4	GA	JR	JR, GA
JR	40:60	67.1	55	4			
EC	40:60	64.2	10111	8			
DC	20:80	65.7	7602	7			

From table 3, general conclusion can be made based on dengue dataset. GA is the best model in term of *acc* and shorter rl while JA is good at generating lower *nr* as well as shorter *rl*. The mathematical formulas define the best model among reduct method and are written as follows:

- $acc\ (R_{GA} \rangle R_{JA,} R_{EC,} R_{DC}) \rightarrow b_{M}\_acc(R_{GA})$
- $nr\ (R_{JA} \langle R_{GA} , R_{EC,} R_{DC}) \rightarrow b_{M}\_rl(R_{GA})$
- $rl\ (R_{JA}, R_{GA} \langle R_{EC,} R_{DC}) \rightarrow b_{M}\_nr(R_{JA,} R_{GA})$

CONCLUSION

The analyzed rule can be seen as knowledge to be used in future intelligent decision application. Following that, classification system will be considered as intelligence if it has ability to produce lesser and shorter knowledge but capable to classify more unseen cases with high accuracy result [5]. Thus, the selection of proper classification techniques for dengue is an important task.

In this paper, four reduct algorithm in RS are evaluated. The aimed of evaluation is to search for the best reduct knowledge generator for dengue dataset. The comparative study is carried out for reduct algorithm namely GA, JR, EC, and DR in term of *acc, nr, rl,* and *cov*. The experimental result shows that the performances of all methods are comparable at *acc* and for several method, it obviously generate huge *nr* and *rl*. GA seems to be the most accurate classifier with shorter rule. As good as GA, JA method also produces comparable *acc* but distinctly generate lower and shorter rule. For EC and DR, both of them capable to generate high score of *acc* but use more *nr* and longer rules to obtain similar *acc*. Beside that, EC and DR require longer time during training even though the number of *tn* dataset is low. In conclusion, JA is chosen as the best method for knowledge acquisition for dengue dataset. The selection is made based on the capability of JA to obtain good *acc* from limited number of rule and shorter rule.

REFERENCES

[1] B. G. Buchnan and E. H. Shortliffe, *Rule-based expert system*. New York: Addoson-Wisely, 1984.

[2] T. Shusaku, "Automated extraction of hierarchical decision rules from clinical databases using rough set model," *Expert System with Application*, vol. 24, pp. 189-197, 2003.

[3] G. Holmes and S. J. Cunningham, "Using data mining to support the construction and maintenance of expert systems," *Artificial Neural Networks and Expert Systems*, pp. 156-159, 1993.

[4] J. Han and M. Kamber, *Data Mining: Concept and Technique*, 2 ed. New York: Morgan Kaufman, 2006.

[5] A. A. Bakar, "Propositional Satisfiability Method in Rough Classification Modelling For Data Mining." vol. Phd Serdag, Malaysia: University Putra Malaysia, 2002.

[6] M. Holsheimer, M. Kersten, M. Mannila, and H. Toivonen, "A Perspective on Databases and Data Mining," CWI. Netherlands. 1995.

[7] T. W. Miller, *Data and Text Mining: A Business Application Approach*. New Jersey: Pearson Education, 2005.

[8] J. Wroblewski, "Finding Mininal Reduct Using Genetic Algorithm," in *Proceeding of the 2nd Annual Join Conference on Information Science*, Wrightsville Beach, NC, 1995, pp. 186-189.

[9] A. Ohrm, *ROSETTA technical reference manual*. Trondheim, Norway: Norweigan University of Science Technology, 1999.

[10] J. G. Bazan, "Dynamic Reduct and Statistical Inference," in *Information Processing and Management of Uncertainty in Knowledge Based System (IPMU'96)*, Granada, Spain, 1996.

[11] Z. Pawlak, "Rough sets and data analysis," in *Fuzzy Systems Symposium, 1996. 'Soft Computing in Intelligent Systems and Information Processing'., Proceedings of the 1996 Asian*, 1996, pp. 1 - 6

[12] W. Ziarko, "Discovery through rough set theory " *Communications of the ACM*, vol. 42, pp. 54-57, 1999.

[13] H. S. Nguyen, "Descretization problem for rough set methods," in *In Proc of First Int. Conf. on Rough Set and Current Trend in Computing*, 1998, pp. 545-552.