Data Mining Reduction Methods and Performances of Rules

Faudziah Ahmad[a], Mohammad Aizat Basir[b]

[a]*College of Arts and Sciences*
Universiti Utara Malaysia, 06010 UUM Sintok, Kedah Darulaman
Tel : 09-6683314, Fax: 04-9284753
Email : fudz@uum.edu.my

[b]*Faculty of Science and Technology*
Universiti Malaysia Terengganu, Malaysia, 21030 Kuala Terengganu, Terengganu,
Tel : 09-6683314
Email : aizat@umt.edu.my

ABSTRACT

In data mining the accuracy of models are associated with the strength of the rules. However, most machine learning techniques produce a large number of rules. The consequence is with large number of rules generated, processing time is much longer. This study examines rules of different lengths of attributes in terms of performance based on percentage of accuracy. The research adopts the Knowledge Discovery in Databases "KDD" methodology for analysis and applies various data mining techniques in the experiments. Data of 50 hardware dataset companies which, contains 31 attributes and 400 records have been used. In summary, results show that in terms of performance of rules, Genetic Algorithm has produced the highest number of rules followed by Johnson's Algorithm and Holte's 1R. The best classifier for extracting rules in this study is VOT (Voting of Object Tracking). In terms of performance of rules, best results comes from rules with 30 attributes, followed by rules with 1 intersection attribute and lastly rules with 3 intersection attributes. Among the three sets of attributes, the 3 intersection attributes are considered as the attributes that can be used as predictor attributes.

Keywords
Reduction, Rough Set, Companies performance, Rule extraction, Knowledge Discovery in Databases

1.0 INTRODUCTION

Data Mining is the important element in the more universal process of Knowledge Discovery in Databases (KDD) where its function is to extract useful knowledge from large quantities of data. Data mining has become trendy in many fields due to the growth in data warehouses and the realization that the massive operational data compiled over the years in organizations has the potential to be exploited using intelligent techniques. Port (2001) defines data mining as a tool that harnesses artificial intelligence and statistical measurements to uncover hidden knowledge. Data mining has been made for extensive analysis to spot subtle relationships and associations as well as uncovers patterns in data using predictive techniques. Reduction is an important process in data mining where its purpose is to diminish the level redundancy of knowledge that can be represented as decision rules. Rules are patterns that have been uncovered by data mining process and play a critical role in decision makings as decisions can be made easier and faster. The process of generating rules is known as rules extraction. Specifically, rule extraction is a description of the hypothesis that is comprehensible yet closely approximates the data's prediction behavior.

There are many types of rules and some examples are propositional if-then rules, M-of-N rules, oblique rules, equation rules, fuzzy rules, and first order logic rules.

Rule extraction can be evaluated by number of criteria (Craven and Shavlik, 1999):
a. Quality of the extracted rules – involved an aspect of rule quality such as accuracy, fidelity, comprehensibility.
b. Scalability of the algorithm – Scalability refers to how the running time of a rule-extraction algorithm and the comprehensibility of its extracted models vary as a function of such factors as the underlying model, the size of the training set and the number of input features c. Consistency of the algorithm – consistency as the ability of an algorithm to extract rules with the same degree of accuracy under different training sessions [1].

The most common criteria used in evaluating rules is accuracy. Accuracy is probably much more important than consistency in data mining application areas. In other domains, a different situation may occur where consistency plays a more important role. In data mining, the deliverable is a model. In genaral, a model uses an algorithm to act on a set of data. The notion of automatic discovery refers to the execution of data mining models. Data mining models can be used to mine the data on which they are built, but most types of models are generalizable to new data. The process of applying a model to new data is known as testing or model evaluation.

Models generated can be measured using rules, percentage of accuracy, support and confidence level. For measuring rules, strong rules are those with the highest confidence value (Mennis & Liu, 2005). Rules of less than 30% are considered as weak rules. Good models are associated with strong rules while models that are not categorized as 'good' are associated with weak rules. In data mining, several techniques can be used to produce rules. Examples are the Neural Networks, Fuzzy Logic, Genetic Algorithms (GA), Johnson's Algorithm, Holte's 1R, Decision Tree and Apriori. However the numbers of rules produced by these techniques are usually enormous. As a result, the processing time is longer. Performing reduction on a set of data is one mechanism to decrease the number of rules. The reason being reduction process will eliminate attributes that are irrelevant and thus rules will be generated from the reduced number of attributes. There are many methods for performing reduction. These are GA, Johnson, Holte's 1R and others. These methods are unique in their own way and produce reduced set of attributes. However, the question is which reduced set of attributes should be used to generate a model? Thus, a research is required to analyze the performances of various reduction methods.

In this study three reduction methods namely GA, Johnson, and Holte's 1R has been experimented. The models generated by these methods are evaluated using percentage of accuracy. Specifically, the study intends to investigate the performance of the three reduction methods, and identify the best reduction method for the set of used.

2.0 RELATED WORKS

Companies' performance can be analyzed using data mining techniques such as Neural Networks, Fuzzy Logic, Genetic Algorithms, Statistics, Rough Sets, Stepwise Regression, Decision Tree and Association Rules.

Kappert & Omta (1997) applied neural network to recognize pattern or relationship. They concluded that neural networks provide some interesting features when modeling real business data. Neural network has been proven to be a viable option for data mining tasks and management and can be used to extract knowledge from vast amounts of operational data. Besides neural network, fuzzy logic has been claimed to be able to represent unstructured knowledge. Zadeh (1965) claimed that a fuzzy set was a natural choice for knowledge representation since the prediction involves imprecise concepts and imprecise reasoning. Many research findings so far indicated that fuzzy logic and sets are complementary. In a similar study, the combination of fuzzy logic and rough sets was investigated by Lin (1998). Besides fuzzy logic, genetic algorithms have also been applied with rough set. Gargano, Marose & Kleeck (1991) applied genetic algorithms (GA) for investment analysis. GA concepts incorporated for the investment analysis include survival of the fittest, crossover and mutation. GA process involves the maximization of a set of investment alternatives with the objective of maximizing the annual return subject to a set of constraints. According to Gargano, the difficulty with using GA in investment analysis was in the function and inequalities that were not well-suited to the traditional programming languages and methodologies. GA required an initial population of feasible points, an evaluation function, conventions for creating potential new members of the population by mating or random mutation, and a grim reaper mechanism to delete poorly performing members in order to make room for ones.

In the area of business and finance, rough set theory have been explored to solve business problems by several researchers. Puagwatana & Gunawardana (2005) developed a model to predict business failure in Thailand particular in technology industry by using four variables from Altman's model and adding one variable to the model. The model has been developed by using the stepwise logistic regression. Stepwise logistic regression has been applied to develop a new model from which some variables came from Altman. In the end of their research, they concluded that stepwise logistic regression model have the ability to assist management for predicting corporate problems early enough to avoid financial difficulties. Moreover, the evidence from analysis of warning signs from the model can signal going concern problems early before eventually enters bankruptcy. In another research work, Segovia-Vargas, Gil-Fana & Heras-Martínez (2003) proposed an approach to predict insolvency of insurance companies based on Rough Set Theory. In the end of the experiment, they claimed that Rough Set Theory is a competitive alternative to existing bankruptcy prediction models in insurance sector and have great potential capacities that undoubtedly make it attractive for application to the field of business classification.

Rough set has also been explored in transporation area. For example, Sawicki, Zak & Wlodarczak (2003) presented the concept of quality evaluation of the transportation system by means of the Rough sets theory. They found that rough set theory has a potential to determine the set of decision rules that are useful in the quality evaluation of a transportation system.

From the researches highlighted above, rule extraction has been found to be an important process to extract knowledge and can be used as a basis to identify important factors that influence businesses. Wang, Xu & Wang (2004) discovered that rough set can be used to extract and filter rules. Li and Cercone (2005) conducted a research related to rule extraction and discovered the potential of association rules to give a better performance. Another research that has been done by Jun & Song (2006) showed that fuzzy decision trees can provide a small number of rules, simple structure of the tree and high classification accuracy.

In summary all researches highlighted above show the importance of rules extraction. Rules with good classification accuracies are highly required to produce reliable knowledge. This is important as good decision makings are based on these knowledge.

3.0 METHODOLOGY

The study adopts GMDR (General Methodology of Design Research) and KDD (Knowledge Discovery in Databases) approach. This methodology has been proposed by Vaishnavi and Kuecheler (2004). The steps are shown below:

3.1 Problem Identification

This phase includes establishing the problem of the research. The objectives, scope and significance of the study are also identified.

3.2 Requirement Gathering

This phase includes activities such as requirements gathering and data collections. This phase also includes finding the appropriate techniques for conducting data mining process.

3.3 Rule Extraction

This is the main phase where the KDD Process is applied. Several experiments have been conducted on various lengths of rules. During this stage several sub processes have been conducted. Brief explanation of the processes is given below:

a) Data Selection

The data used through out this study has been obtained from previous research. The initial data contains factors that influence the survivability of hardware companies. The total number of records experimented was 400 with 31 financial attributes from over 50 hardware companies in Malaysia. Examples of attributes are Current Asset (CA), Current Liability (CL), Work Cost (WC) and Total Asset (TA). All the data contains numeric value. The target is represented as Net Income (NI). Details of the attributes can be found in Faudziah (2006).

b) Preprocessing Data

This step includes two subsection, data transformation and handling missing values and noisy data. Since data used in this study is in numeric form, it is thus not necessary to do the data transformation task. The data also does not have missing values or noisy data as it has been previously used in another research.

c) Discretization

Discretization process involves converting continuous values into categories or classes. Examples of discretization techniques are Equal Frequency Binning, Boolean Reasoning Algorithm and Entropy/MDL Algorithm. Not all techniques are suitable to discretize the dataset used in this study. In this experiment, Equal Frequency Binning has been used for discretization as the technique has been found to produce the best results. The end result of this process is data are transformed into several categories.

d) Split Data

In this step data is divided into two folds, train data and test data. The trained data set is reduced and rules are generated from this data. The test data has been used to verify the rule generated from train data. Split factor 0.2 is used as this technique has been found to produce high accuracy from previous experiment.

e) Reduction

Data reduction is a process to diminish the level of redundancy of knowledge that can be represented as decision rules or classification. Data reduction is one of the steps in Rough Set Theory and is used to compute the minimal attribute or reduct of attribute in the databases. It has been known that the use of a set of attributes (reduct) without loss of any essential information is better than the use of the entire set of attributes. The database can be reduced by removing attributes which are considered as not important. Several reduction techniques are Genetic Algorithm (GA), Johnson's Algorithm and Holte's 1R. GA has been used in this experiment because from previous experiment, it has been found to produce the best classification accuracy. Miller and Thomson (2003) claimed that GA are better for interpreting the feature space since they consistently identify groups of variables that yield better results. They also found GA be able to increase the specificity of the classifier while maintaining the sensitivity.

f) Data mining

In this step, a suitable data mining task has been identified. The examples of data mining tasks are such as clustering, classification, association and others. For this research, classification task has been carried out. Several classification techniques such as Standard voting (SV), Voting with Object Tracking (VOT), Naïve Bayes (NB) and Standard / Tuned Voting (STV) have been tested. Voting with Object Tracking (VOT) has been found to be the most appropriate technique for this research. Results obtained (percentage of accuracies) from VOT have been better than other techniques of classification. Ten fold cross-validation a method has been applied in all experiments for validation purposes. The output for this phase is a set of core attributes that is capable to produce a good prediction model.

3.4 Evaluation

Results produced have been evaluated based on percentage accuracy. Rules containing attributes with the highest percentage of accuracy has been selected.

3.5 Conclusion

Several experiments have been conducted to test different size of reducts and reducts with the highest accuracy are selected.

4.0 RESULTS AND FINDINGS

Experiment has been conducted using 3 different reduction methods. Table 1 below shows the summary of results obtained:

Table 1: *Number of Rules Generated from 3 Reduction Method*

Reduction Method	**Num of Rules Generated**
Genetic Algorithm	8091
Johnson's Algorithm	274
Holte's 1R	150

Table 1 shows the number of rules generated from 3 different reduction methods. The result shows that Genetic Algorithm has produced the highest number of rules followed by Johnson's Algorithm and Holte's 1R. For this reason, GA has been chosen for further experiment because it produced highest number set of rules that can be analyzed to extract important features.

Table 2: *Reduct Lengths Obtained from GA*

Num	Selected Attributes (Reduct)	Length
1	{MVE, EBIT, FV6, EPS}	4
7	{SE, EBIT, FV8, BVPS, EPS}	5
30	{BVTD, SP, EBIT, FV6, ROA, PrcB}	6
110	{BVTD, Sho, RE, EBIT, AltB, BVPS, EPS}	7

The total number (NUM) of set obtained from Genetic Algorithm is 1, 7, 30 and 110 reducts. The lengths of reducts obtained are 4, 5, 6, and 7. The length of reducts also indicated the number of attributes contained in the rules. Further experiments have been conducted to test different lengths of reducts. Split factor 0.2 and Voting with Object Tracking (VOT) classifier has been used in the experiments. Table 1.3 below shows the result for length 4, 5, 6, and 7.

Table 3: *Summary of Result for Reduct Length 4, 5, 6 and 7*

Length	Average of accuracy from 10 randomly selected data
4	0.599498
5	0.576059
6	0.501136
7	0.477728

From the Table 3 it can be seen that when different lengths of reducts have been applied to ten randomly selected data, results obtained are below than 60%. This indicates that the attributes in the reducts have little influence on the datasets. For this reason, further experiments have been conducted to identify the influential attributes. This has been done by looking into the intersection of selected attributes in length 4, 5, 6, and 7. Table 4 shows the results of using the intersection attributes. Also, the results show comparisons of intersections attributes and attributes that have not been reduced.

Table 4: *Rules with 30 attributes and rules with intersection attributes*

30 attributes and Intersection attributes	Average of accuracy from 10 randomly selected data
30 Attributes	0.936745
Intersection of Length 4 [MVE, EBIT, EPS]	0.807942
Intersection of Length 4, 5, 6 and 7 [EBIT]	0.925038

From Table 4 it can be seen that reducts with 30 attributes gives 94% percentage of accuracy. Reducts with three intersection attributes give 81% of accuracy and reduct with one intersection attribute give 93% of accuracy. Based on the results it can be seen that rules with 30 attributes give the best results, followed by reducts with one attribute and reducts with 3 attributes. However, 30 attributes is too many to be used for modeling data and although it has shown to produce highest results, the use of all attributes is not efficient in terms of processing time, data collection and operating costs.
The use of one attribute, on the other hand, although saves in terms of processing time, data collection and cost, but do not really show the reliability of the results as it is doubtful that 1 attribute could be used to model the whole data set.

The use of 3 attributes although produce the least percentage of accuracy, can still be considered as good since the percentage of accuracy obtained is more than 70%. In summary it can be said that 30 attributes and 1 intersection attribute are 2 extreme cases and the use of the attributes are not considered as appropriate. However, rules with 3 intersection attributes are chosen as the good set of rules because these attributes have achieved minimum requirement of percentage of accuracy for good models.

5.0 CONCLUSION

The study attempts to show the capability of a data mining technique known as rough set theory to extract useful knowledge. Rough Set Theory has demonstrated that important features and rules can be extracted to predict the survivability of hardware companies.

The study has successfully achieved all three objectives. In terms of performance of rules, Genetic Algorithm has produced the highest number of rules followed by Johnson's Algorithm and Holte's 1R. Next, the best classifier for extracting rules in this study is VOT (Voting of Object Tracking). In terms of performance of rules, best results comes from rules with 30 attributes, followed by rules with 1 intersection attribute and lastly rules with 3 intersection attributes. However, among the three sets of attributes, the 3 intersection attributes are considered as the attributes that can be used as predictor attributes.

Due to time constraint the study is limited to the use of three rule extraction techniques namely GA, Johnson, and Holte's 1R. Future enhancement can be conducted to improve the findings. Some suggestions are:

- Use other techniques (algorithms) in discretization, reduction and classification on the same kind of problem. It may produce better result or new findings.
- Add new factors (attributes) as new factors may have strong influence in predicting the survivability of hardware companies.

The research shows a method to identify a good set of predictor on companies' success or failure. Thus, the research can assists companies' stakeholders such as investment analysts, financial analyst, bankers, managers, and others to identify performing and non performing companies.

REFERENCES

Craven M.W. and Shavlik J.W. (1999). Rule extraction: Where do we go from here?

Deboek, G.J. (1995)"Trading on the Edge: Neural, Genetic and Fuzzy Systems for Chaotic Financial Markets", John Wiley & Sons.

Faudziah A. (2006). Penentuan Indikator Ketahanan Syarikat E-Dagang Menggunakan Pendekatan Set Kasar, Ph.D Tesis, Universiti Kebangsaan Malaysia.

Jeremy Mennis, Jun Wei Liu, 2005. Mining Association Rules in Spatio-Temporal Data. *Transactions in GIS*. 9(1): pp. 5-17

Jun, Y.B. and Song, S.Z. (2006). Generalized fuzzy interior ideals in semigroups. *Information Science*, 176(20): 3079-3093.

Gargano, M.L., Marose, R.A. and Kleeck, L.V. (1991)
"An application of artificial neural networks and genetic algorithms to personnel selection in the financial industry," in *Proc. 1st Int.Conf. Artificial Intelligence on Wall Street*. Los Alamitos, CA:
IEEE Computer Soc., 257–262.

Kappert, C.B. and Omta, S.W.F. (1997). Neural networks in technology management processes, in *Information Systems V*, IEEE Computer Society Press, California, 465-473.

Li, J. and Cercone, N. (2005). Discovering and ranking important rules. In *IEEE Granular Computing*, volume 2, pages 506–511, Beijing, China, July 2005.

Lin, T.Y. (1998). "Granular Computing on Binary Relations I: Data Mining and Neighborhood Systems." *In: Rough Sets In Knowledge Discovery*, A. Skowron and L. Polkowski (eds), Physica-Verlag, 107-121.

Miller, J. and Thomson, P. (2003). Cartesian genetic programming. In Riccardo Poli, Wolfgang Banzhaf, William B. Langdon, Julian F. Miller, Peter Nordin, and Terence C. Fogarty, editors, *Third European Conference on Genetic Programming*, volume 1802 of *Lecture Notes in Computer Science*. Springer, 2000.

Port, O. (2001). New tools are revealing the secrets hidden in mountains of data, *The Business Week* 50, 2001, Issue 3726A, pp. 185-188.

Puagwatana, S.; Gunawardana,, K.D. (2005). Business failure prediction model: a case study of technology industry in Thailand, Engineering Management Conference, Proceedings. IEEE International Volume 1, Issue , Sept. 11-13, 246 – 249.

Sawicki, P., Zak, J., and Wlodarczak, H. (2003). Rough Sets Based Quality Evaluation of the Road Freight Transport System.

Segovia-Vargas M.J., Gil-Fana J.A., Heras-Martínez A., Vilar-Zanón, J.L. and Sanchis-Arellano A. "Using Rough Sets to predict insolvency of Spanish non life insurance companies". 2003.

Vaishnavi V. and W. Kuechler (2004). "Design Research in Information Systems" . Retrieved from
URL: http://www.isworld.org/Researchdesign/ drisISworld.htm.

Wang, X., Xu, R. and Wang, W. (2004). Rough Set Theory: Application in Electronic Commerce Data Mining, *Web Intelligence*, 541-544

Zadeh, L A. (1965). Fuzzy sets. *Inform. Contr.* **8**:338-53, Dept. Electrical Engineering and Electronics Res. Lab, Univ. California, Berkeley, CA.