Investigating the optimal number of attributes to manage knowledge performances

Mohammad Aizat b. Basir, *Faculty of Science and Technology, Universiti Malaysia Terengganu, Malaysia,*
aizat@umt.edu.my
Faudziah bt. Ahmad, Faculty of Information and Technology, Universiti Utara Malaysia, Malaysia,
fudz@uum.edu.my

Abstract

Rules are the most important element in knowledge extraction. The performance or strength of rules will determine how good a model is. Higher accuracy implies that a model is good and vise versa. However, the strength of rules depends on the attributes. The number of attributes in a rule can influence the percentage of accuracy in a model. 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 investigated the performances of rules with different lengths of attribute and identified the optimal number of rule for a good model. The research performed experiments using several data mining techniques. Data of 50 hardware dataset companies which, contains 31 attributes and 400 records was used. Results showed that in terms of number of rules, Genetic Algorithm (GA) 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 set with 3 attributes are considered as the best and three (3) has been identified as the optimal number of attributes.

1. Introduction

Data Mining is method that has been widely used to extract knowledge. The main idea of data mining is to extract some useful knowledge from large quantities of data. Data mining has become trendy in many fields especially in business IT. The emergence is due to the growth in data warehouses and the realization that this mass of operational data has the potential to be exploited as an extension of Business Intelligence. [1] defines data mining as a tool that harnesses artificial intelligence and slick statistical tricks to unearth insights hiding inside mountains of data. It has also been made for extensive analysis and the ability to spot subtle relationships and associations, that it regularly makes fresh discoveries. Data mining uncovers patterns in data using predictive techniques. These patterns play a critical role in knowledge discovery. The patterns uncovered using data mining form an archive of knowledge for organizations. Thus, investigating the patterns via rules and its attributes can help organizations choose appropriate set of knowledge for managing their business operations.

Rule form the basic foundation of knowledge extracted from databases. In machine learning there are many types of rules. These are:-

I. Propositional if-then rules – the condition part of a propositional rule is a Boolean combination of conditions on the input variables.
II. M-of-N rules – closely related to propositional rules.
III. Oblique rules – represent piecewise discriminant functions.
IV. Equation rules – more complex than Oblique rules and contains a polynomial equation in the condition part.
V. Fuzzy rules – the rules are expressed in terms of linguistic concepts and corresponding membership functions.
VI. First Order Logic – rules that can contain quantifiers and variables.

Rule extraction can be evaluated by number of criteria:-

I. Quality of the extracted rules – involved an aspect of rule quality such as accuracy, fidelity, comprehensibility.
II. 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 (Craven and Shavlik, 1999).

III. Consistency of the algorithm – consistency as the ability of an algorithm to extract rules with the same degree of accuracy under different training sessions [2].

Accuracy is probably much more important than consistency in data mining application areas. In other domains a different situation may occur in which consistency plays an important role when expounding power is the main requirement.

In data mining, the accuracy of models generated can be measured using rules, percentage of accuracy, support and confidence level. For measuring rules, the strong rules are those with the highest confidence value [3]. On the other hand, rules of < 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, GAs, 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. Beside this, all the rules will be executed even though some may not be relevant. There is thus, a need to find a method to obtain a good set of relevant rules. This research intends to extract rules obtained by rough set technique and identify a good set of relevant rules.

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 results to an increase in operating cost, and labor. On the other hand, short rules too have some weaknesses. Though processing time is much shorter, the accuracy of models developed from short rules are sometimes incompetence. Poor percentages of accuracies are one of the results that have been cited. Thus, a research is required to analyze the performances of various lengths of rules so that the best length of rules can be presented. As such, this study attempts:-

I. To investigate the performance of rules where generated using GA, Johnson's Algorithm and Holte's 1R.

II. To identify the best classifier for rules extraction.

III. To investigate the performance of rules with various length of attributes.

2. Related Works

Companies' performance can be analyzed using data mining techniques such as Neural Networks, Fuzzy Logic, GAs, Statistics, Rough Sets, Stepwise Regression, Decision Tree and Association Rules. The following session will elaborate some of the methods.

2.1 Neural Networks

[4] in their paper tiltled "Neural Networks and Business Modeling - An Application of Neural Modeling Techniques to Prospect Profiling in the Telecommunications Industry" from University of Groningen used neural network to recognize pattern or relationship in the analysis. They concluded that neural networks provide some interesting features when considering the difficulties encountered in modeling real business data. They also mentioned that neural nets has proven to be a viable option for data mining tasks and management as the technique can be used to extract knowledge from vast amounts of operational data.

2.2 Fuzzy Logic

[5] claimed that a fuzzy set was a natural choice for knowledge representation since the prediction involves imprecise concepts and imprecise reasoning. He defines a fuzzy trading system as consisting for one or more inputs, a fuzzy rule base, and one or more output variables. The input are fuzzified, membership functions, and fuzzy associations are defined between the inputs and outputs. The fuzzy outputs derived from the system was defuzzified into trading recommendations. 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 [6].

2.3 GA

[7] applied GA for investment analysis. GA concepts incorporated for the investment analysis included survival of the fittest, crossover and mutation. GA process involves maximization of a set of investment alternatives with the objective of maximizing the annual return, which is, subjected to a

set of constraints. Giordano pointed out that the difficulty with these types of problems has been in the function and inequalities that were not well-suited to the traditional programming languages and methodologies. GA requires an initial population of feasible points, an evaluation function, conventions for creating potential new members of the population, and a grim reaper mechanism to delete poorly performing members.

2.4 Stepwise Regression

[8] developed a model to predict business failure in Thailand particularly in the technology industry by using four variables from Altman's model and a new variable. The model was developed by using the stepwise logistic regression. Their results concluded that the stepwise logistic regression model have the ability to assist management in predicting corporate problems early enough to avoid financial difficulties. Moreover, the evidence from analysis of warning signs can signal going concern problems earlier before eventually falling into bankruptcy.

2.5 Rough Set

[9] presented the concept of quality evaluation of the transportation system by means of the Rough Sets theory. Rough Set theory has a potential to determine the set of decision rules that are useful in the quality evaluation of a transportation system [2]. [10] proposed an approach to predict insolvency of insurance companies based on Rough Set Theory. Their results showed that Rough Set Theory is a competitive alternative to existing bankruptcy prediction models in insurance sector and have great potentials in the field of business.

From the researches that have been mentioned above, rule extraction has been found to be important in extracting knowledge and has been used in business to help people identify important factors that influence businesses.

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. 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 (2005). The steps are shown below:-

3.1.1 Awareness of Problem

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

3.1.2 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.

3.1.3 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 explanations of the processes are given below:

a) Data Selection

The data used through out this study has been obtained from previous research conducted by [11]. The initial data contains factors that influence the survivability of 50 Hardware Companies in Malaysia. The total number of cases is 400 with 31 attributes. Examples of attributes are Current Asset (CA), Current Liability (CL), Work Cost (WC) and Total Asset (TA). All data are in numerical form.

b) Preprocessing Data

This step includes two subsections, data transformation, and handling missing values and noisy data. The data does not have missing values or noise as it has been previously used in another research.

c) Discretization

This step is the most critical part as it can affect the overall experimental results. Here, continuous values are changed into classes. This task has been done by dividing the range of the attribute into classes or categories (1, 2, 3, 4 or 5). The categories are then used to replace the actual data values. On completing this step, all data are in the form of category 1, 2, 3, 4, or 5. A Rough Set software known as ROSETTA has been used to perform the discretization. In ROSETTA, several discretization techniques can be used. These are Equal Frequency Binning, Boolean Reasoning and Entropy/MDL Algorithm. Since not all techniques are suitable, several experiments have been conducted to identify the best method for discretizing the dataset. From the experimental results obtained, Equal Frequency Binning has been found to produce the best result and thus has been chosen for the discretization.

d) Split Data

In this step data was divided into two sets, that is train data and test data. Split factor 0.2 was used to divide the dataset. This technique was applied as it was found to produce high accuracy from previous experiment. Train data was used for generating rules while the test data was used to verify the accuracy of the rules generated.

e) Reduction

Data reduction is a process to remove 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 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. In the Rough Set, several reduction techniques are GA, Johnson's Algorithm and Holte's 1R. However, GA has been used in this experiment because from previous experiment it has produced the best classification accuracy. Miller et al. (2003) has mentioned that GA are better for interpreting the feature space since they consistently find groups of variables that yield better results. They also found that GA is 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 clustering, classification, and association. 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 method has been applied in all experiments for validation purposes. The output for this phase is a set of core attributes known as 'reduct" that is capable of producing a good prediction model.

3.1.4 Evaluation

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

Several experiments have been conducted to test various lengths of rules. The study identified which length of rules produce the best result.

7. Finding and Results

Experiment has been conducted using 3 different reduction methods. The table below shows the summary of results obtained:-

Table 1.1: Number of Rules Generated from 3 different Reduction Method

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

Table 1.1 shows the number of rules generated from 3 different reduction methods. The result shows that GA 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 1.2: Rule Length 4, 5, 6, and 7 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 of set (Num) obtained from GA is 110 reducts. The lengths of rules obtained are 4, 5, 6, and 7. The length of rules also indicated the number of attributes contained in the rules. For further experiment, all 110 reduct will be tested based on the rule length. For this experiment, split factor 0.2 and Voting with Object Tracking (VOT) classifier has been used. Table 1.3 below shows the result for length 4, 5, 6, and 7.

Table 1.3: Summary of Result for Length 4, 5, 6 and 7

Length	Average of accuracy from 10 randomly selected data
4	0.599498 (60%)
5	0.576059 (58%)
6	0.501136 (50%)
7	0.477728 (48%)

From the table 1.3 it can be seen that when different lengths of rules have been applied to ten randomly selected data, results obtained are below than 60% (0.599498 *100). This indicates that the attributes in the rules do not have influence on the data sets. 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 1.4 shows the results of using the intersection attributes. Also, the results show comparisons of intersections attributes and attributes that have not been reduced.

Table 1.4: 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 (94%)
Intersection of Length 4 [MVE, EBIT, EPS]	0.807942 (81%)
Intersection of Length 4, 5, 6 and 7 [EBIT]	0.925038 (93%)

From Table 1.4 it can be seen that rules with 30 attributes gives 94% percentage of accuracy. Rules with three intersection attributes give 81% of accuracy and rules 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, following rules with one attribute and rules 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.

8. Significance and Contributions

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.

9. 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, GA 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.

10. References

[1] PORT, O., Spring, Virtual ProspectingFrom oil exploration to neurosurgery, new tools are revealing the secrets hidden in mountains of data, *The Business Week* 50, 2001, Issue
3726A, pp. 185-188.

[2] Neumann. J. Classification and evaluation of algorithms for rule extraction from artificial neural networks, 1998.

[3] Johansson, U., Niklasson, L. and Konig, R. Accuracy vs. Comprehensibility in Data Mining Models.
Retriebed 23 May, 2004, from
http://www.fusion2004.foi.se/papers/IF04-0295.pdf.

[4] C.B. Kappert et al., "Neural Nerworks and Business Modelling-An Application of Neural Modelling Techniques to Prospect Profiling in the Telecommunications Industry", System Sciences, Jan. 1997, pp. 465-473.

[5] Zadeh, L., Fuzzy sets, Information Control 8, 1965, pp. 338-353,

[6] Lin, T.Y. "Granular Computing on Binary Relations II: Rough Set Representations and Belief Functions." In: Rough Sets In Knowledge Discovery, A. Skoworn and L. Polkowski (eds), Springer-Verlag, 1998, 121-140.

[7] Gargano, M. L. and Raggad P. "Data Mining – A Powerful Information Creating Tool".
Retrieved May 23, 2008 from
http://www.pafis.shh.fi/graduates/supsir01.pdf

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

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

[10] 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.

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