Indicator Selection based on Rough Set Theory

Faudziah Ahmad[a], Azuraliza Abu Bakar[b], Abdul Razak Hamdan[c]

[a]College of Arts and Sciences
Universiti Utara Malaysia, 06010 UUM Sintok, Kedah, Malaysia
Tel: 04-9284787, Fax : 04-9284753
Email : fudz@uum.edu.my

[b]Faculty of Technology and Information Science
Universiti Kebangsaan Malaysia, 46300 Bangi, Selangor, Malaysia
Tel: 03-89216794, Fax : 03-89256732
Email : aab@ftsm.ukm.my

[c]Faculty of Technology and Information Science
Universiti Kebangsaan Malaysia, 46300 Bangi, Selangor, Malaysia
Tel: 03-89216342, Fax : 03-89256732
Email : arh@ftsm.ukm.my

ABSTRACT

A method for indicator selection is proposed in this paper. The method, which adopts the General Methodology and Design Research approach, consists of four steps: Problem Identification, Requirement Gathering, Indicator Extraction, and Evaluation. Rough Set approach also has been applied in the Indicator Extraction phase. This phase consists of 5 steps: Data selection, Data Preprocessing, Discretization, Split Data, Reduction, and Classification. . A dataset of 427 records have been used for experimentation. The datasets which contains financial information from several companies consists of 30 dependant indicators and one independent indicator. The selection of indicators is based on rough set theory where sets of reducts are computed from a dataset. Based on the sets of reducts, indicators have been ranked and selected based on certain set of criteria. Indicators have been ranked through computation of frequencies in reduct sets. The major contribution of this work is the extraction method for identifying reduced indicators. Results obtained have shown competitive accuracies in classifying new cases, thus showing that the quality of knowledge is maintained through the use of a reduced set of indicators.

Keywords*:*
companies' performance, indicators selection, reduction, extraction, rough set

1.0 INTRODUCTION

Financial indicators have been found to have great influence on organizational performance. Financial data has been a popular source and used to analyze companies' performance by many research companies such as Multex Investor, Media General Financial Services Corporate, Nasdaq,and Reuters. Financial analysis, which focuses on financial information, in general, can be categorized into profitability ratio, efficiency ratio, and price ratio. Measures from these categories are many and among these are current ratio, quick ratio, net income, working capital, operational income, revenue, sales growth, earnings per share, gross profit, book value, stock price, stock volume, and others (Corrado & Jordon, 2000). These measures have been found to have great influence on the performance or companies and are indicators of the success of companies. All these measures are relevant indicators in measuring success. However, to include all relevant indicators in the measurement would present a tremendous burden in terms of data collection, analysis, and cost. Evidence in the literature, indicates that there are a limited number of critical areas necessary to the successfulness functioning of organizations (Rockhart, 1979). Three indicators has been commonly used (Globerson , Globerson & Frampton, 1991), however, not more than seven indicators has been recommended (Globerson, 1985). The use of rough set has been explored in various financial areas such as prediction of business failure (Dimitras, Slowinski, Susmaga & Zopounidis, 1999), stock market analysis (Golan & Ziarko 1995; Grzymala-Busse, 1997; Tay & Shen, 2002), and marketing (Beynon, Curry & Morgan, 2000). In terms of feature selection, Hu & Shi (2003) have proposed a novel feature ranking technique using

discernibility matrix. Dash & Liu (1997) have applied Rough Set technique on a feature selection problem to obtain patterns of customers and products. By making use of indicators' information in the discernibility matrix, a fast feature ranking mechanism has been produced.

This paper explores the use of rough set to select indicators. A fundamental problem in a company's information system is whether the whole information available is always necessary to represent the success of the company. Using the concept of reduct in rough set theory, the set of interesting indicators is determined (Pawlak, 1991). Interesting indicators are indicators that are more important than the rest of indicators within the database. Computation through reducts has been performed to obtain a set of indicators. This paper makes use of reduct computation techniques to rank and identify a set of interesting indicators. It also attempts to show the relevance of using a smaller set of indicators as compared to a larger one. The paper is organized as follows. The next section, focused on proposed method. Section 3 and 4 present the experimental results and conclusion respectively.

2.0 INDICATOR SELECTION METHOD

In order to ensure the research meets the requirements, the General Methodology of Design Research (GMDR) has been used throughout the study. The methodology is shown in Figure 1.

2.1 Problem Identification

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

2.2 Requirement Gathering

Activities such as requirements gathering and data collections are performed during this phase. Information have been obtained through interviews and materials from books, journals, companies' reports, companies newsletters, and other documents from the Internet.

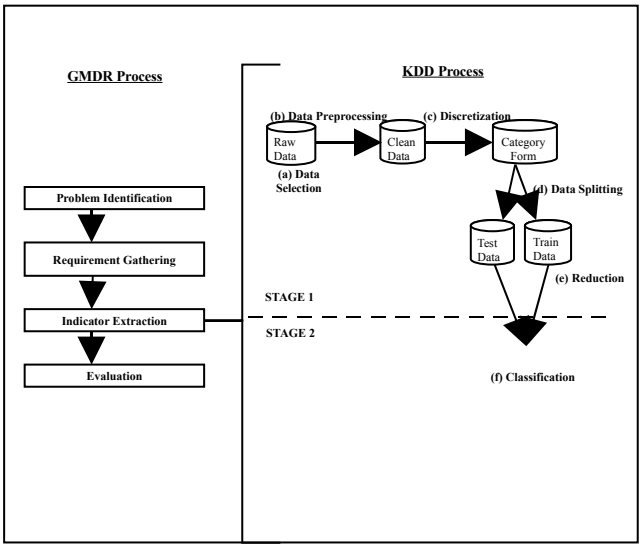

Figure 1: Indicator selection method

2.3 Indicator Extraction

This is the main phase where the KDD Process has been applied. Several experiments have been conducted to choose suitable data mining techniques. Some brief explanations of the processes are given below:

a) Data Selection
Data used throughout this study has been obtained from previous research. The initial data contains factors that influence the survivability of companies. The total number of records is 427 and each record contains 31 attributes of financial indicators. Some examples are Current Asset (CA), Current Liability (CL), Work Cost (WC) and Total Asset (TA). The target attribute is Net Income (NI) and all data are in numeric form. Table 1 shows sample of the original data. Details of the indicators can be obtained from Faudziah , Azuraliza, and Abdul Razak (2005).

Table 1: *Sample of original data*

Num	CA	CL	WC	.. .	NI
1	105783.00	19374.00	86409.00	...	7716.00
2	92397.00	20594.00	71803.00	...	8177.00
3	172644.00	40978.00	131666.00		6148.00
4	234659.00	50982.00	183677.00	...	2848.00
5	323257.00	90387.00	232870.00	...	27200.00
6	391327.00	86062.00	305265.00	...	20821.75
7	479040.00	91440.00	387600.00	...	41803.00
8	570251.00	96731.00	473520.00	...	55897.00
...	...	...	...	...	...
425	6126000.00	1934000.00	4192000.00	...	-86000.0000
426	6122000.00	1678000.00	4444000.00	...	117000.0000
427	6447000.00	1641000.00	4806000.00	...	121000.0000

b) Data Preprocessing
This step includes handling missing values and noisy data. Missing values have been replaced with the average values. Noisy data have been identified using box plot and scatter diagrams.

c) Discretization
In this step, continuous values are changed into classes. This step is the most critical part and has been taken seriously as discretized data could greatly affect the performance of the model generated and thus, affects the end result of the study. There are several techniques that can be used to discretize data. Examples are Equal Frequency Binning (EFB), Boolean Reasoning (BR), Entropy (ENT), Naïve (NV), Semi-Naïve (SNV) and manual. Each techniques of discretization has been developed to cater certain problems and thus may not be suitable to use in all circumstances. Discretization techniques that have been tested are BR, ENT, EFB, NV , SNV, and manual cuts. Manual cuts have been done by dividing the range of the attribute values into intervals. Interval labels were then used to replace actual data values. After discretization, data values are represented with several classes. The discretization using BR, ENT, EFB, NV, and SNV have been done using Rosetta (Rough Set Technical Analysis Software). Manual cuts have been conducted manually and are based on statistical calculation. Figure 2 shows the process of choosing the best discretization technique:-

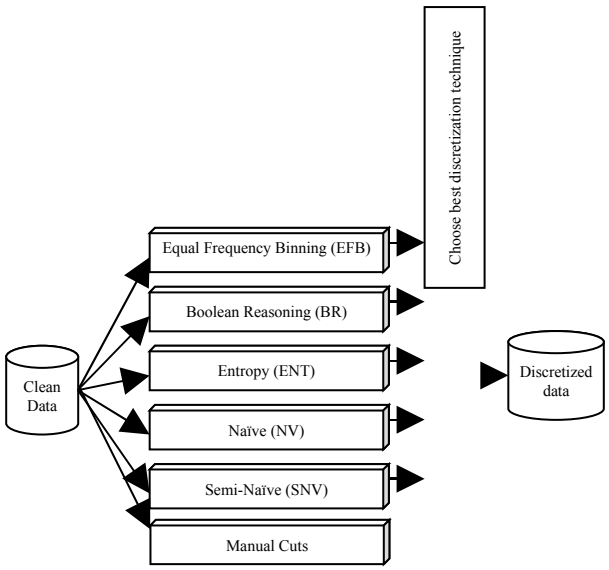

Figure 2: Process of choosing the discretization technique

Table 2 shows sample of the discretized data.

Table 2: *Sample of discretized data*

Num	CA	CL	WC	TA	SE	...	NI
1	1	1	1	1	1	...	2
2	1	1	1	1	1	...	2
3	1	1	1	1	1	...	2
4	1	1	1	1	1	...	2
…	…	…	…	…	…	…	…

d) Data Splitting
In this step, data is divided into two sets, train data and test data using several splitting techniques. The splitting techniques or also known as split factor that have been experimented are 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 and 0.95. Splitting technique 0.2 denotes that 20 % of data are allocated for training and 80% are for testing. The reason for performing experiments on 10 different splitting techniques is to identify the best split technique. Using the best split technique in the Indicator Extraction phase could contribute to getting a good model. The train data has been used to get a model while the test data has been used to verify the model. The model is evaluated in terms of percentage of accuracy, number of rules, and percentage of errors. Table 3 below shows the split factor and data divisions used for training and testing. Figure 3 shows the process of choosing the split factor technique.

Table 3: *Splitting technique and data division*

Split Factor	Train Data	Test Data
0.1 (10%)	10% (43 objects)	90% (384 objects)
0.2 (20%)	20% (85 objects)	80% (342 objects)
0.3 (30%)	30% (128 objects)	70% (299 objects)
0.4 (40%)	40% (171 objects)	60% (256 objects)
0.5 (50%)	50% (214 objects)	50% (213 objects)
0.6 (60%)	60% (256 objects)	40% (171 objects)
0.7 (70%)	70% (299 objects)	30% (128 objects)
0.8 (80%)	80% (342 objects)	20% (85 objects)
0.9 (90%)	90% (384 objects)	10% (43 objects)
0.95 (95%)	95% (406 objects)	5% (21 objects)

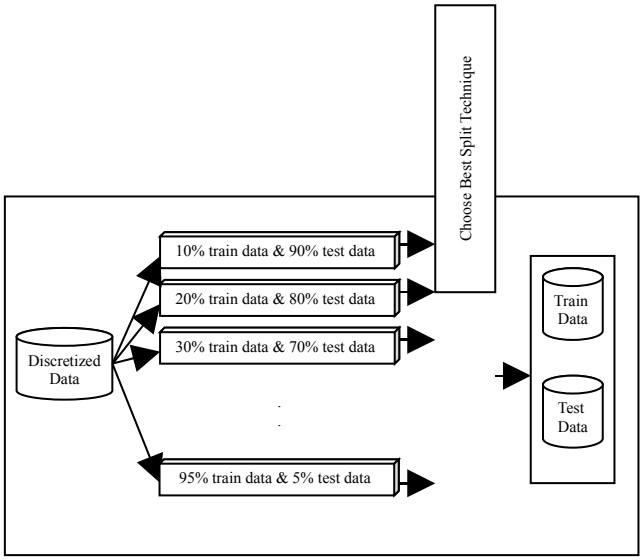

Figure 3: Process of choosing the split techniques

e) Reduction

Reduction is a process to eliminate the redundancy of knowledge. It is one of the steps in Rough Set Theory and is conducted by computing the minimal attributes required. Minimal attributes is also known as "reduct of attributes" in Rough Set Theory. The use of a reduced set of attributes without loss of any essential information has been known to be better than the use of the entire set of attributes. For achieving better performance in generating a model, a database can thus, eliminate unimportant knowledge or redundancy by removing attributes which are considered as not important. There are several reduction techniques and among these are Genetic Algorithm (GA) (Vinterbo and Ohrn, 2000), Johnson's Algorithm (Johnson, 1974) and Holte's 1R (Holte, 1993). These three techniques have been experimented on the discretized dataset and results comprised of several sets of reducts. The reduction method that gave the highest accuracy have been chosen as the best reduction technique to be used in the study. Figure 4 shows the process of performing reduction.

INPUT: Financial data
OUTPUT Accuracy (%)
****

1. *Start*
2. *Read data*
3. *Use-technique[3]={ GA, Johnson, Holtes1R };*
4. *i = Use-technique*
5. *Divide data using chosen split technique*
6. *Set round = 1*
7. *While (round < 4) do*
   - 7.1 *Extract reduct using i*
   - 7.2 *Test reduct*
   - 7.3 *Record result*
   - 7.4 *Next i*

   End

Figure 4. Reduction process

f) Classification

In this step, several classification techniques such as Standard voting (SV), Voting with Object Tracking (VOT), Naïve Bayes (NV) and Standard / Tuned Voting (STV) have been tested. The classification accuracy more than 70% has been considered as good classification techniques. Classification technique that gives highest average of accuracy has been chosen as the best classification techniques. The outcome of this process is the best classification technique. Figure 5 shows the process of choosing the best classification technique:

INPUT: Financial data
OUTPUT Accuracy (%)
****

1. *Start*
2. *Read EDT data*
3. *Class-Algo = SV, VOT, NB, STV*
4. *i = Class-Algo*
5. *Divide data using chosen split technique*
6. *Reduce data using chosen reduction method*
7. *Set round = 1*
8. *While (round less than 5) do*
   - 8.1 *Perform classification using i*
   - 8.2 *Record result*
   - 8.3 *Next i*

   end

Figure 5. Classification process

2.4 Evaluation

Results obtained from experimenting sets of attributes with different length have been analyzed in terms of percentage accuracy. The set of attributes that obtained the highest percentage of accuracy has been selected as the attributes that provide the best set of rules.

3.0 RESULTS AND ANALYSIS

In this study 427 records of financial data have been experimented. Each record contains 30 indicators (independent) and 1 dependent indicator. Several experiments have been conducted following the steps highlighted in section 2(b) till 2(f). The experiments conducted were aimed to choose a suitable discretization technique, split technique, reduction method, and classification technique. Based on the experimental results, the chosen techniques were Equal Frequency Binning (discretization technique), 0.2 SF (split factor technique), GA (reduction technique), and SV (classification technique). These techniques were then used to extract important indicators from the dataset. It was found that nine indicators have been selected. These are BVPS(90%), EBIT(90%), EPS(80%), FV3(60%), MVE(50%), ROA(50%), ROE(50%), SHO(60%), and SpgC(50%). These indicators have been identified as important indicators that have some influence on the companies performance.

In order to verify the indicators, ten-fold cross validation tests have been conducted. Ten-fold cross validation technique is used to ensure the consistency of results. The results are shown in Table 5 below.

Table 4. *Results of Ten-fold cross validation tests*

After Reduction			
FOLD	*ACC (%)*	*NR*	*ERROR (%)*
1	85	682	15
2	91	1212	9
3	91	689	9
4	84	2140	16
5	84	840	16
6	82	380	18
7	67	1315	33
8	76	485	24
9	73	952	27
10	74	735	26
Average	80.7		19.3

The table shows classification accuracy (*ACC*), number of rules generated (*NR*), and percentage of error (ERROR). The best model comes from the fold that has the highest % accuracy. From the table, it has been found that FOLD 2 and 3 produced the same percentage of accuracy i.e 91% with 1212, and 689 respectively. This shows that the smaller amount of rules or knowledge can produce a good decision and represent the whole information system well. The average accuracy of 80.7% indicates that these nine indicators are essential to represent knowledge and model the success of an EC company. In addition, the results show that there is a good blend of data in each dataset. Ten folds validation technique is used to handle the fluctuation of accuracies. The average accuracies in all dataset are above 70% indicating that the model is acceptable for further consideration.

The results are promising as the best model in each category is able to correctly classify other respective categories with more 70% accuracy. It indicates that these limited numbers of indicators and rules are crucial in all categories of dataset in giving accurate decisions.

4.0 CONCLUSION

In this paper, a method for indicator selection has been proposed. The study has been conducted following the General Methodology and Design Research approach. This approach consists of four steps: Problem Identification, Requirement Gathering, Indicator Extraction, and Evaluation. Rough Set approach also has been applied in the the Indicator Extraction phase. This approach consists of 6 steps: Data selection, Data Preprocessing, Discretization, Split Data, Reduction, and Classification. These steps were constructed based on the idea of reduct computation and feature ranking in the theory of rough set. The results are measured in terms of percentage of accuracy, number of rules and percentage of errors.

A dataset of 427 records have been used for experimentation. The experimental results showed that out of 30 indicators, nine have been found to be adequate in representing the whole knowledge of the dataset. These indicators when tested for validity using ten-fold cross validation method showed good accuracies. Although several folds showed fluctuation, the average percentage of decreased in accuracy in each dataset was not significant. Thus, this indicated that the volume of knowledge after reduction is adequate to make a decision. This study attempted to assist companies in deciding which indicators to focus from a whole group of indicators.

REFERENCES

Beynon, M., Curry, B., and Morgan, P., (2000). Classification and rule induction using rough set theory. *Expert Systems*, Vol.17, No.3, 136-147.

Corrado, C.J. and Jordon, B.D. (2000). *Fundamentals of Investment: Valuation and Management*, Boston: McGraw-Hill International Editions.

Dash, M., and Liu, H., (1997). Feature selection for classification, *Intelligent Data Analysis*, Vol. 1, 131-156.

Dimitras, A.I., Slowinski, R., Susmaga, R. and . Zopounidis, C. (1999). Business failure prediction using rough sets. *European Journal of Operational Research*, 114, 263-280.

Faudziah A., Azuraliza A.B., Abdul Razak H. (2005). "Rough Approach for Information Synergy: Case Studies on E-Commerce Companies", *Proceedings of the International Conference of Information and Communication in Management*, 23-25 May, Melaka, Malaysia.

Johnson, D.S. (1974). Approximation Algorithms for Combinatorial Problems. *Journal of Computing System Science*, 9(3), 256-278.

Globerson, S. (1985). Issues in Developing a Performance Criteria System for an Organisation. *International Journal of Production Research,* 23(4), 639-646.

Globerson, A., Globerson, S. and Frampton, J. (1991). *You can't Manage What You Don't Measure*, Aldershot UK: Avebury.

Golan, R.H., and Ziarko, W. (1995). A Methodology for Stock Market Analysis Utilizing Rough Set Theory. *Proceedings of Computational Intelligence for Financial Engineering*.

Greco, S., Matarazzo, B. and Slowinski, R. (1998). A new rough set approach to multicriteria and multiattribute classification. *in L. Polkowski and A. Skowron, Rough Sets and Current Trends in Computing, Springer-Verlag, Berlin*, 60-67.

Grzymala-Busse, J.W. (1997). A new version of the rule induction system LERS, Fundamenta Informaticae, v.31 n.1, 27-39.

Holte, R.C. (1993). "Very Simple Classification Rules Perform Well on Most Commonly Used Datasets", Machine Learning, 11, 63-92.

Hu, K., Lu, Y. and Shi, C. (2003). Feature ranking in rough sets. AI Commun, 16(1), 41-50.

Kauffman, R. J. and Wang, B. (2001). *The Success and Failure of Dotcoms: A Multi-Method Survival Analysis*. Retrieved May 2008, from http://misrc.umn.edu/wpaper/WorkingPapers/kw_cist2001_submission_92501.doc.

Meyer, P. and Pifer, H. (1970). Prediction of Bank Failures, *Journal of Finance*, 853-868.

Ohlson, J. (1980). Financial ratio and the Probablistic Prediction of Bankruptcy, *Journal of Accounting research*, Vol 18, 109-131.

Ohrn, A. (1999). Discernibility and Rough Sets in Medicine: Tools and Applications. Department of Computer and Information Science, Norwegian University of Science and Technology, Trondheim, Norway.

Ohrn, A. 2001. Rosetta Technical Reference Manual, Knowledge System Group, Norwegian University of Science and Technology, Trondheim, Norway.

Pawlak, Z. (1991). Rough Sets: Theoretical aspects of reasoning about data, *System Theory, Knowledge Engineering and Problem Solving*, Dordrecht: Kluwer.

Rockart, J.F. (1979). Chief executives define their own data needs, *Harvard Business Review*, 30, 81-92

Tay, F. E. H. and Shen, L. (2002). Economic and financial prediction using rough sets model, *European Journal of Operational Research*, 141.

Shirata, C. Y. (2003). *Financial Ratios as Predictors of Bankruptcy in Japan: An Empirical Researc.* Retrieved May 2008, from http://www3.bus.osaka-cu.ac.jp/apira98/archives/pdfs/31.pdf

Vinterbo, S. and Ohrn, A. (2000). Minimal Approximate Hitting Sets and Rule Templates, *International Journal of Approximate Reasoning* **25**(2), 123-143.