Classifying Firms' Performance using Data Mining Approaches

Bahtiar Jamili Zaini[#1], Massudi Mahmuddin[*2]

[#]*School of Quantitative Sciences, Universiti Utara Malaysia,*
06010 UUM Sintok, Kedah, Malaysia
[1]bahtiar@uum.edu.my
[*]*School, of Computing, Universiti Utara Malaysia,*
06010 UUM Sintok, Kedah, Malaysia
[2]ady@uum.edu.my

***Abstract*— Superior prediction and classification in determining company's performance are major concern for practitioners and academic research in providing useful or important information to the shareholders and potential investors for investment decision. Generally, the normal practice to analysed firm's performance are based on financial indicators reported in the company's annual report including the balance sheet, income and cash flow statements. In this work, a few popular and important benchmarking machine learning techniques for the data mining including neural networks, support vector machine, rough set theory, discriminant analysis, logistic regression, decision table, sequential minimal optimization and decision tree have been tested as to classify firm's performance. The data mining techniques produce high classification rate that is more than 92%. This work also has reduced total number of ratios to be evaluated due to long processing time and large processing resources. Finally, the CA/TA, S/TA, E/TA, GM, FC, PBT/TA, and EPS have been considered for of the final reduced financial ratios. The results show that the 7 reduced ratios are comparable as the common 24 ratios. And to the still produce high classification rate and able classify the firm's performance.**

***Keywords*—** *Feature Selection, Financial Ratios, Classification, Data Mining*

1. Introduction

Abundant of studies investigated on the prediction of the stock's return and firms' performance using these financial reports [1] – [2]. Financial statement analysis is useful for investors and management of the firm because it can be used to help them to predict and classify the firms' performance which is gave higher income in terms of future earnings and dividends. From management's standpoint, financial statement analysis is useful both to help anticipate future conditions and as a starting point for planning actions that will improve the firm's future performance [3].

Firm's performance can be analysed based on financial indicators reported in company's annual report; balance sheet, and income statement. These reports are the indicators on how they operated their firms and managed their financial resources. These reports also provide vast amount information of the firm's performance. To provide a better understanding on the performance of the particular firm, these financial data are best transformed into financial ratios. Financial statement analysis is useful for the investors and management to predict and classify future earnings and dividends.

However, there are a lot of financial ratios to be considered in classifying the performance of each firm. Some of these financial ratios may be irrelevant and correlate among them, so could give redundant information for classification. In addition, there are no standard financial ratios use to determine the performance of the firms in many studies and each firm will use different ratios analysing firm's performance. Hence, using all the ratios to build the classification model, then, the system will operate in a high dimension. This computationally not viable and analytically to be complicated [4]. In addition, regression method and neural networks of data mining techniques are difficult to use when the number of features (features, indicators and financial ratios will be interchangeably employed in this work) are large [5]. Therefore, discovering the optimal financial ratios is very important because it effects on the accuracy and classification of the developed model.

In reducing data dimensionality, feature selection can be employed. It is one of the important steps in

International Journal of Supply Chain Management
IJSCM, ISSN: 2050-7399 (Online), 2051-3771 (Print)

data mining process. A goal of feature selection is to avoid selecting too many or too few features than is necessary. Feature selection is a process that selects a subset of original features, and reduces the number of features, removes irrelevant, redundant, or noisy data, and brings the immediate effects for applications: speeding up a data mining algorithm, improving mining performance such as predictive accuracy and result comprehensibility [6]. Hence, feature selection is an important step in most of data mining problems, including predicting, since good performance of prediction models comes from enhanced sorted data [7].

This study uses the common financial ratios that are widely used by the firm's in bankruptcy prediction. Based on selected financial ratios, models were built to classify firm's performance using various classification techniques such as neural networks, rough set theory, decision tree, support vector machine, sequential minimal optimization (SMO), decision table, discriminant analysis, and logistic regression. Then, the best optimal or subset of financial ratios using feature selection techniques is identified. The comparison between all ratios included in the models and just using a subset of ratios in classifying firm's performance were compared.

The remaining part of this paper is organized as follows. Section 2 summarizes the literature review on the financial ratios and prediction model used on firm's performance. Section 3 explained the methodologies used in this study. Then section 4 shows the results and analysis of the research. Finally, section 5 concludes the papers.

2. Literature Review

Financial indicators are very important tool to describe and analyse the business operation performance. Those indicators or ratios play an important role to evaluate and forecast the company operation performance and financial situation. Researchers and economists evaluate and analyse firm's performance using variety of financial ratios. Abundant of studies investigated on the prediction of the stock's return and firms' performance using financial ratios. There are many different financial ratios used to analyse firms' performance. Different studies used different financial indicators in analysing firm's performance. Therefore, it is important to discover the most important financial ratios as it will affect the accuracy and classification of the model developed. The differences methodologies between researchers not only differ in terms of financial ratios used, but also the techniques used to determine optimal ratios and also classifier used to classify firm's performances. Kumar and Ravi [8] gave a comprehensive review on previous work from 1968 to 2005 in solving the bankruptcy prediction problem. They indicated that neural networks are the most widely applied technique in bankruptcy prediction followed by statistical model, rough sets, case-based reasoning, operational research, and other techniques. Other researchers reviewed the techniques used in bankruptcy prediction, such as rough set [9], and neural networks [10].

Although the rough set theory has been rarely used to build the model for predicting bankruptcy shows the best performance among them [7], [11] – [14]. Slowniski *et al.* [11] compared rough set approach with discriminant analysis for prediction of company acquisition in Greece. Dimitras *et al.* [12] compared rough set with logit analysis and discriminant analysis to predict business failure in Greece. Zhou *et al.* [13] showed that integrated hybrid method with rough set and support vector machine, and rough set with back propagation neural networks outperform with discriminant analysis, back propagation neural networks and support vector machine to the problem of credit risk assessment. Ahn *et al.* [14] used rough set and hybrid with neural network showed the proposed technique outperform with discriminant analysis and neural network alone. Paik and Suh [7] found that when predicted delisting firms in Korea, the hit ratio of rough set is the highest compared with other techniques, followed by decision tree, support vector machine, linear regression, and neural network.

Feature selection is one of the important steps in data mining process. Feature selection technique has become very actively researched with applications in varieties of fields for it is one of the most important issues in research fields such as data mining and pattern recognition. The goal of feature selection to data reduction, decrease noise, and eliminate irrelevant or redundant features. Many researchers have used varieties of feature selection methods in financial prediction with the objective to improve the classification accuracy. Because financial firms' performance prediction

involves many financial ratios to be considered, therefore feature selection is an important step in pre-processing data. A few works [12] – [16] used rough set approach to reduce the number of financial ratios. Wang and Chen [15] used 11 financial ratios and one decision attribute which was classifying company's performance into good, medium, and bad and applying rough set theory to corporate credit ratings. Dimitras *et al.* [12] compared rough set with logit analysis and discriminant analysis to predict business failure in Greece and used 28 financial ratios. Bose [16] used rough set theory to evaluate dot-coms firms' financial health using 24 financial ratios. Paik and Suh [7] used 66 financial ratios when predicting delisting firms in Korea. They used rough set theory, decision tree, support vector machine, linear regression, and neural network. Their results showed that rough set get highest classification rate compared with others. Slowinski *et al.* [11] compared rough set approach to discriminant analysis to predict company acquisition in Greece. Ahn *et al.* [14] used 8 ratios and hybrid rough set with neural network. Zhou *et al.* [13] used 12 financial ratios and integrated hybrid method with rough set and support vector machine and rough set with back propagation neural networks for the problem of credit risk assessment. Olson and Mossman [17] used 61 financial ratios and neural network to forecast Canadian stock return. Comparing their result to ordinary least squares and logistic regression, they concluded that neural network outperformed both techniques. Hua *et al.* [18] used 22 financial ratios and support vector machine applying to the problem of bankruptcy prediction. Wu *et al.* [19] proposed hybrid genetic algorithm with support vector machine and used 19 financial ratios to predict bankruptcy in Taiwan. They compared the result to discriminant analysis, logistic regression, probit regression, neural network and by using support vector machine alone and found that the proposed methods gave the highest predictive accuracy. Ko and Lin [20] used particle swarm optimization, genetic algorithm, and stepwise statistical analysis to reduce 39 financial ratios then employing linear regression, discriminant analysis, and neural network to forecast financial distress in Taiwan. They found that integrating more techniques achieve better forecasting. Lin and McClean [21] used variety data mining approaches such as discriminant analysis, logistic regression, neural network, and decision tree to predict corporate failure in United Kingdom. Among the individual classifiers, decision tree and neural networks were found to provide better performances. They also hybrid several techniques and produce better results than individual classifiers. Karbhari and Sori [22] used 64 ratios and discriminant analysis to predict corporate financial distress. Bose [16] used rough set theory to evaluate dot-coms firms' financial health using 24 financial ratios. Wang and Chen [15] used 11 financial ratios and one decision attribute which was classifying company's performance into good, medium, and bad and applying rough set theory to corporate credit ratings.

3. Methodology

3. 1 Data Collection

This study used financial statements data of firms listed in Bursa Malaysia Main Board from 2001 until 2010. The financial ratios are calculated by exploited the financial indicators information in firm's balance sheet and income statement as published in the annual report. There were 24 financial ratios to be considered to this study. These 24 ratios were adapted from study conducted by [23]. The performance of firms in the stock market can be classify into two categories, high performing and low performing based on their return on equity. The firms are classified as high performance if its return on equity is in the top 25% of this ranking. Meanwhile, the lowest 25% of return on equity for each year are classified as low performance. According to [24], the 25% cut-off is selected so as to give a clear difference between the mean returns on the two groups. Therefore, for each year, all the firms were sorted according to their return on equity of their firm. Table 1 list the numbers of firms which is classified as high performance and low performance of firms for each year. Table 2 lists the financial ratios used for this study.

Table 1. The total number of firms for each type of performance

Year	Low (0)	High (1)	Total
2001	85	85	170
2002	90	90	180
2003	107	107	214
2004	132	132	264
2005	132	132	264

2006	115	115	230
2007	141	141	282
2008	145	145	290
2009	133	133	255
2010	150	150	300
Total	**1230**	**1230**	**2460**

Table 2. List of financial ratios used for this study

Financial Ratio	**Terms**
Current Assets/Current Liabilities	CA/CL
Current Assets/Sales	CA/S
Current Assets/Total Assets	CA/TA
Current Liabilities/Total Assets	CL/TA
Working Capital Ratio	WCR
Total debt/Total Assets	TD/TA
Debt/equity	D/E
Interest Expenses/Sales	IE/S
Current Liabilities/Shareholders' Equity	CL/SE
Sales/Total Assets	S/TA
Earnings/Sales	E/S
Earnings/Total Assets	E/TA
Earnings/Current Liabilities	E/CL
Gross Margin	GM
Net Tangible Asset Backing per share	NTAS
Liquid Asset per share	LAS
Free Cash flow to capital	FC
Earnings/Gross Profit	E/GP
Gross Profit/Total Assets	GP/TA
Profit before Taxes/Total Assets	PBT/TA
Profit before Taxes/Current Liabilities	PBT/CL
Profit before Taxes/Sales	PBT/S
Dividend per share	DPS
Earnings per share	EPS

Furthermore, the dataset was divided into two datasets, one for training data and another for testing data. The training data used to train and generate models while the testing data used to classify the firms and validated the performance. There were about 70% of the data used for training data and 30% for testing data. Then, ten samples were derived from dataset using a random seed.

3.2 Feature Selection and Classifying

In this stage, neural networks, support vector machine, rough set theory, discriminant analysis, logistic regression, decision table, SMO and decision tree are applied to classify firm's performance. These methods were widely used in bankruptcy prediction [8-10]. Performance of various classification techniques are compared based on classification rate.

In order to identify the optimal financial ratios, variety of feature selection techniques were conducted so that only these optimal financial ratios are used to classify firm's performance. Several feature selection techniques such as best first, exhaustive search, genetic search, greedy stepwise, linear forward selection, random search, rank search, scatter search, and subset forward selection to reduce the number of financial ratios are employed. The aim for conducting a several feature selection techniques are to identify the optimal financial ratios based on reducing the number of financial ratios. This subset of financial ratios, when classify firm's performance should give high classification rate. Out of 24 financial ratios, the number of financial ratios are reduced and then used this subset of ratios to conduct again the experiment to classify firm's performance. The results of classification rate using all ratios and the reduced set of ratios were compared. Figure 1 shows the flowchart of the methodology for this study.

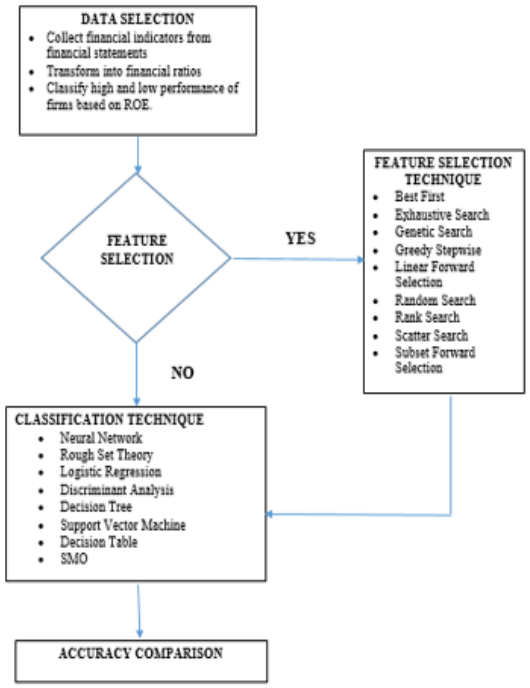

Figure 1. The flowchart of the Methodology

4. Analysis and Finding

4.1 Performance of Classification Techniques

In this step, performance of various data mining models when classifying firm's performance are compared. Again, neural networks, support vector machine, rough set theory, discriminant analysis, logistic regression, decision table, sequential minimal optimization (SMO) and decision tree are applied to classify firm's performance. Performance of these models is presented in Table 3.

Table 3. Classification rate for different methods

Methods	Classification Rate (%)
Logistic Regression	98.90
Decision Table	98.90
Neural Network	98.82
Rough Sets Theory	98.67
Decision Tree	98.58
SMO	97.52
Discriminant Analysis	95.20
Support Vector Machine	92.76

The results show that all these methods could give high classification rate which is almost greater than 92% correct classification rate. Logistic regression and decision table models show the highest classification rate among the eight models. The support vector machine shows the lowers classification rate which is 92.76% but still considered as good results. The good model's classification rate achieved in this study implies that those models have a very good ability to classify firm's performance. These results suggest that using all 24 financial ratios could give a good result in classifying firm's performance.

4.2 Reducing the Number of Financial Ratios

The aim for this step is to reduce the number of financial ratios so that based on this subset of financial ratios still give high classification rate when classify firm's performance. Out of 24 financial ratios, the reduced subset of original ratios is identified using several feature selection techniques. The purpose conducting the feature selection is to reduce the number of ratios while maintaining acceptable classification accuracy. This study uses several feature selection techniques and find which ratios appeared mostly in these feature selection techniques.

Based on Table 4, ratios CA/TA, S/TA, E/TA, GM, FC, PBT/TA, and EPS were found to occur most frequently in almost of these feature selection techniques. Therefore, for the next of experiment, only these seven financial ratios are used to classify the high and low performance of firms.

4.3 Classifying Firms' Performance Using a Reduced Sets of Financial Ratios

Using the same financial dataset but just the reduced set of financial ratios instead of all 24 ratios, the experiment to classify firm's performance is again conducted. Table 5 gives the comparison results of classification rate using all ratios and the reduced set of ratios.

Table 5. Classification rate for different methods

Methods	All Ratios (%)	Reduced Ratios (%)	Differences (%)
Logistic Regression	98.90	98.05	0.86
Decision Table	98.90	98.37	0.53
Neural Network	98.82	97.76	1.07
Rough Sets Theory	98.67	94.24	4.34
Decision Tree	98.58	97.80	0.78
SMO	97.52	96.30	1.25
Discriminant Analysis	95.20	94.00	1.20
Support Vector Machine	92.76	91.42	1.45

This result shows that although using the reduced set of financial ratios to classify firms' performance still can give high classification rate. Based on the reduced set of ratios, decision table models give highest classification rate which is 98.37%. This classification rate less 0.53% as compared to the whole financial ratios. The support vector machine modals give the lowest classification rate compared to others eight models. The classification rate of 91.42% still give a good

Table 4. Analysis on feature selection techniques

Financial Ratio	Best First	Exhaustive Search	Genetic Search	Greedy Stepwise	Linear Forward Selection	Random Search	Rank Search	Scatter Search	Subset Forward Selection	AVERAGE %
CA/CL	0	0	10	0	0	0	0	0	0	1.11
CA/S	0	0	10	0	0	0	0	0	0	1.11
CA/TA	**40**	**40**	**10**	**40**	**40**	**40**	**40**	**40**	**40**	**36.67**
CL/TA	0	0	0	0	0	0	0	0	0	0.00
WCR	10	10	10	10	10	10	10	10	10	10.00
TD/TA	0	0	0	0	0	0	0	0	0	0.00
D/E	0	0	10	0	0	0	0	0	0	1.11
IE/S	0	0	30	0	0	0	0	0	0	3.33
CL/SE	0	0	0	0	0	0	0	0	0	0.00
S/TA	**50**	**50**	**10**	**50**	**50**	**50**	**50**	**50**	**50**	**45.56**
E/S	0	0	80	0	0	0	0	0	0	8.89
E/TA	**10**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**90.00**
E/CL	0	0	80	0	0	0	0	0	0	8.89
GM	**90**	**90**	**0**	**90**	**90**	**10**	**90**	**90**	**90**	**71.11**
NTAS	0	0	0	0	0	0	0	0	0	0.00
LAS	0	0	20	0	0	0	0	0	0	2.22
FC	**10**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**90.00**
E/GP	0	0	100	0	0	90	0	0	0	21.11
GP/TA	0	0	60	0	0	0	0	0	0	6.67
PBT/TA	**10**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**90.00**
PBT/CL	0	0	30	0	0	0	0	0	0	3.33
PBT/S	0	0	70	0	0	10	0	0	0	8.89
DPS	0	0	30	0	0	0	0	0	0	3.33
EPS	**10**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**90.00**

results and just lost 1.45% compared to whole financial ratios. Meanwhile the rough set theory gives the biggest differences classification rate when compared using all ratios and reduced ratios. This finding suggests that the reduced set of ratios such as CA/TA, S/TA, E/TA, GM, FC, PBT/TA, and EPS can be used to classify firm's performance and still give good results.

5. Conclusion

In this study, several data mining techniques to classify firm's performance are used based on their characteristic financial statements. There were 24 financial ratios are being considered with high and low firm's performance. All these data mining techniques gave high classification rate more than 92%. The number of original ratios then reduced and identified the optimal ratios using several feature selection techniques. CA/TA, S/TA, E/TA, GM, FC, PBT/TA, and EPS are identified as the reduced subset of financial ratios. Based on these seven ratios, the firm's performance again is classified using several data mining techniques and the results still gave high classification rate. Therefore, by using these reduced ratios, also could classify firm's performance and produce as comparable results.

Acknowledgments

This research is partly supported by RAGS scheme no. 12689, Universiti Utara Malaysia.

References

[1] J. L. Gissle, D. Giacomino, and M. D. Akers, "A Review of Bankruptcy Prediction Studies: 1930-Present". Journal of Financial Education, Vol. 33, 2007.

[2] A. Raph, "Financial Ratio as an Instrument for Evaluating Company Performance-An Overview". International Journal of Banking, Finance, Management & Development Studies, 3 (25), pp. 430-446, 2015.

[3] E. F. Brigham, and J. F. Houston, *Fundamental of Financial Management Ninth Edition*. Orlando: Harcourt College Publishers, 2001.

[4] M. Pechenizkiy, S. Puuronen, and A. Tsymbal, "The Impact of Sample Reduction on PCA-based Feature Extraction for Supervised Learning". In Proceeding of 21st ACM Symposium on Applied Computing, ACM Press, 2006.

[5] A. Lendasse, J. Lee, E. de Bodt, V. Wertz, and M. Verleysen, "Input Data Reduction for the Prediction of Financial Time Series". In Proceeding European Symposium on Artificial Neural Networks Bruges (Belgium), 237-244, 2001.

[6] H. Liu and L. Yu, "Toward Integrating Feature Selection Algorithms for Classification and Clustering". IEEE Transactions on Knowledge and Data Engineering, Vol. 17, No. 4, 2005.

[7] J. Paik and Y. Suh, "Performance Comparison of Various Data Mining Models to Predict Delisting of Firms". KMIS International Conference, 2005.

[8] P. R. Kumar, and V. Ravi, "Bankruptcy Prediction in Banks and Firms via Statistical and Intelligent Techniques – A Review". European Journal of Operational Research, Vol. 180, 1 – 28, 2007.

[9] F. E. H Tay, and L. Shen, "Economic and Financial Prediction Using Rough Sets Model". European Journal of Operational Research, Vol. 141, 641 – 659, 2002.

[10] A. F. Atiya, "Bankruptcy Prediction for Credit Risk Using Neural Networks: A Survey and New Results". IEEE Transactions on Neural Networks, Vol. 12, No. 14, 2001.

[11] R. Slowinski, C. Zopounidis, and A. I. Dimitras, "Prediction of Company Acquisition in Greece by Means of the Rough Set Approach". European Journal of Operational Research, 100(1), 1-15, 1997.

[12] A. I. Dimitras, R. Slowinski, R. Susmaga, and C. Zopounidis, "Business Failure Prediction Using Rough Sets". European Journal of Operational Research, Vol.114, pp. 263-280, 1999.

[13] J. Zhou, Z. Wu, C. Yang, and Q. Zhao, "The Integrated Methodology of Rough Set Theory and Support Vector Machine for Credit Risk Assessment". IEEE the Computer Society. Proceedings of the Sixth International Conference on Intelligent Systems design and Applications. 2006.

[14] B. S. Ahn, S. S. Cho, and C. Y. Kim, "The Integrated Methodology of Rough Set Theory and Artificial Neural Network for Business Failure Prediction. Expert systems with applications, 18(2), 65-74, 2000.

[15] T.-C. Wang and Y.-H. Chen, "Applying rough sets theory to corporate credit ratings," in Proceedings of the IEEE International Conference on Service Operations and Logistics, and Informatics, pp. 132–136, June 2006.

[16] I. Bose, "Deciding the Financial Health of Dot-Coms Using Rough Sets". Information and Management, Vol. 43, 835 – 846, 2006.

[17] D. Olson, and C. Mossman, "Neural Network Forecasts of Canadian Stock Returns Using Accounting Ratios". International Journal of Forecasting, Vol. 19, 453 – 465, 2003.

[18] Z. Hua, Y. Wang, X. Xu, B. Zhang, and L. Liang, "Predicting Corporate Financial Distress Based on Integration of Support Vector Machine and Logistic Regression". Expert Systems with Applications, Vol. 33, 434 – 440, 2007.

[19] C. Wu, G. Tzeng, Y. Goo and W. Fang, "A Real-Valued Genetic Algorithm to Optimize the Parameters of Support Vector Machine for Predicting Bankruptcy". Expert Systems with Application, Vol. 32, 397 – 408, 2007.

[20] P. C. Ko, and P. C. Lin, "An Evolution-Based Approach with Modularized Evaluations to Forecast Financial Distress". Knowledge-Based Systems, Vol. 19, 84 – 91, 2006.

[21] F. Y. Lin, and S. McClean, "A Data Mining Approach to the Prediction of Corporate Failure". Knowledge-Based Systems, Vol. 14, 189 – 195, 2001.

[22] Y. Karbhari, and M. S. Zulkarnain, "Prediction of Corporate Financial Distress: Evidence from Malaysian Listed Firms during the Asian Financial Crisis". Unpublished Working Paper. Social Science Research Network.

[23] B. J. Zaini, S. M. Shamsuddin, and S. H. Jaaman, "Predicting the Financial Performance of Publicly-Traded Malaysian Firms Using Rough Set Based Feature Selection Techniques". The ICFAI Journal of Applied Finance, 2008.

[24] G. T. Albanis, and R. A. Batchelor, "Predicting High Performance Stocks Using Dimensionality Reduction Techniques Based on Neural Networks". In C. Dunis, and A. Timmerman, editors, Developments in Forecast Combination and Portfolio Choice, pages 117–134. Kluwer Academic Publishers, 2001.