Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong, 19-22 August 2007

FUNDAMENTAL ANALYSIS OF STOCK TRADING SYSTEMS USING CLASSIFICATION TECHNIQUES

CHING-HSUE CHENG, YOU-SHYANG CHEN

Department of Information Management, National Yunlin University of Science and Technology, 123, Section 3, University Road, Touliu, Yunlin 640, Taiwan
E-MAIL: chcheng@yuntech.edu.tw, g9523804@yuntech.edu.tw

Abstract:

The traditional forecasting of revenue growth rate (RGR) is based on normal distribution. Due to emergence of information technology today, data mining has become one of important research trends. Therefore, this paper mainly forecasts revenue growth rate of firms in stock trading systems by classification techniques. It is very important instrument for investors that correctly predict future growing firms from data of fundamental analysis in trading systems, because the accurate prediction of RGR will bring huge profit for investors in the future. This paper proposes a process to predict RGR of firms, which employs Decision tree C4.5, Bayes net, Multilayer perceptron and Rough sets techniques. Moreover, the paper uses the actual RGR dataset in Taiwan stock market to illustrate the proposed process. From the results, we recommend the rough set as analysis tool because the performance is superior to the listing methods and understandable rules are produced.

Keywords:

Revenue Growth Rate; Data Mining Technique; Fundamental Analysis

1. Introduction

In stock trading markets, revenue growth rate is a very important indicator [1] for investors to predict the growing firms in the future. The growing firms stand for particular firm that can ongoing develop and more get gains earnings per share (EPS). Then, investors have been seeking some methods to find out these growing firms with making progress, high EPS and profit in the future. When investors make an investment decision in stock markets, they always may depend on experience or subjective judgment. In general, to prevent this wrong decision-making, there are two kind instruments to aid investors for investment decision-making, which are technical analysis and fundamental analysis. Technical analysis is mainly based on the history data of the correlation between price and volume that reflect investors' behavior. In this paper, we predict revenue growth rate of electronic firm with fundamental analysis [2]. Revenue growth rate is one of core work of fundamental analysis in investment decision-making. Our goal is to find out good prediction tools by fundamental analysis based on data mining techniques. That is, we hope that we can use data mining techniques as tools of forecasting RGR in stock trading systems by finance information of firms.

Data mining, also called Knowledge Discovery in Databases (KDD), is the processing of automatically searching large volumes of digital data for patterns by tools such as classification, cluster analysis, association rules, anomaly detection etc. Data mining is a more complex topic and can link with multiple fields such as computer science and confluence of multiple disciplines such as information retrieval, statistics, databases, algorithms, machine learning and pattern recognition [3, 4]. Clearly, data mining is a field of growing importance for increasing demand for artificial intelligence, fast advance in IT techniques, and processing huge amount of digital data [5, 6]. This study therefore focuses on good forecasting method of classifying revenue growth rate of electronic firms in Taiwan stock trading systems. A new process is proposed to compare and evaluate the results of different techniques in data mining for classifying problems. This process bases on using revenues, assets, profit, income, cost, and other data as condition attributes to determine the potential for future growth of its revenue by four tools of data mining techniques, Decision tree C4.5, Bayes net, Multilayer perceptron and Rough sets.

This paper is organized as follows: Related studies are described in section 2. Section 3 presents the proposed process and practical experimental results; Section 4 is the conclusions of the study.

2. Related work

This section reviews related studies of fundamental

1-4244-0973-X/07/$25.00 ©2007 IEEE

analysis, Decision tree C4.5, Bayes Net, Multilayer Perceptron and the rough set theory.

2.1. Fundamental analysis

Because revenue growth rate is belonging to one of the fundamental analysis, we briefly study the literature for fundamental analysis in this section. One primary objective of fundamental analysis is to explore the relation between financial statement information and future firm fundamental attributes, including revenue growth rate, returns, EPS etc. Fundamental analysis is based on inside environment of firm. The inside environment includes the firm's accounting variable and basic financial status such as price to earning ratio, dividend yield, current ratio, earnings per share, price to book ratio, gross sales, book to market ratio, return on net worth, return on equities, EPS etc [7, 8]. Thus, the objective of this study is to determine whether the fundamental attributes is suit for predicting the revenue growth rate of firms.

2.2. Decision tree C4.5

Classification is an important data mining technique. Many classification models have been proposed, e.g. statistical based, distance based, neural network base and decision tree based [9]. A decision tree is a flow-chart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent class or class distributions [10]. The ID3 [11] is a decision tree algorithm that based on information theory. The basic strategy used by ID3 is to choose splitting attributes with the highest information gain. The concept used to quantify information is called entropy. Entropy is utilized to measure of information in an attribute.

Assume that have a collection set S of c outcomes, then the entropy is defined as

$$H(S)=\sum(-p_i \log_2 p_i) \quad (1)$$

Where p_i the proportion of S is belonging to class i.

$Gain(S,A)$ is information gain of example set S on attribute A is defined as

$$Gain(S,A)=H(S)-\sum\frac{|S_v|}{|S|}H(S_v) \quad (2)$$

And where v is a value of A, S_v = subset of S, $|S_v|$ = number of elements in S_v, $|S|$ = number of elements in S.

C4.5 is a software extension of the basic ID3 algorithm designed by Quinlan [12] to address the following issues not dealt with by ID3: avoiding over fitting the data, determining how deeply to grow a decision tree, reduced error pruning, rule post-pruning, handling continuous attributes, choosing an appropriate attribute selection measure, handling training data with missing attribute values, handling attributes with differing costs, improving computational efficiency.

2.3. Bayes net

A Bayes net is a data structure for fast processing of probability distributions [13]. Bayes nets mainly solve computational problems, and can give exponential reduction in storage. Thus, Bayes net can reduces space complexity of distribution, but does not reduce time complexity for general case. Given the parents, each variable is independent of non-descendents, and Joint probability decomposes [14]. Bayesian net does not necessarily imply a commitment to Bayesian statistics. Bayes net is a useful representation for hierarchical Bayesian models, which form the foundation of applied Bayesian statistics [15, 16].

2.4. Multilayer perceptron

The Multilayer perceptron (MLP) is one of the most widely used types of neural network. It is both simple and solid based on mathematical computation. A typical multilayer perceptron consists of a set of source nodes forming the input layer, one or more hidden layers of computation nodes, and an output layer of nodes [17]. There is always an input layer with a number of neurons equal to the number of variables of the problem, and an output layer, where the perceptron response is made available with a number of neurons equal to the desired number of quantities computed from the inputs [17, 18]. Multilayer perceptron are feedforward neural networks trained with the backpropagation algorithm [18]. They are supervised networks, so they require a desired response to be trained. They learn how to transform input data into a desired response, so they are widely used for pattern classification [19, 20].

2.5. Rough set theory

Rough set is a predictive data mining tool that

incorporates vagueness and uncertainty and can be applied in artificial intelligence and knowledge discovery in databases [21]. Rough set theory, first proposed by Pawlak [22] in 1982, employs mathematical modeling to deal with data classification problems. Rough set addresses the continuing problem of vagueness by applying the concept of equivalence classes to partition training instances according to specified criteria. Two partitions are formed in the mining process. The members of the partition can be formally described by unary set-theoretic operators or by successor functions for upper approximation and lower approximation spaces from which both possible rules and certain rules can be easily derived.

Let $B \subseteq A$ and $X \subseteq U$ be an information system. The set X is approximated using information contained in B by constructing lower and upper approximation sets:

$$\underline{B}X = \{x \mid [x]_B \subseteq X\} \quad \text{(Lower approximation)}$$

And

$$\overline{B}X = \{x \mid [x]_B \bigcap X \neq \varnothing\} \quad \text{(Upper approximation)}$$

The elements in $\underline{B}X$ can be classified as members of X by the knowledge in B. However, the elements in $\overline{B}X$ can be classified as possible members of X by the knowledge in B. The set $BN_B(x) = \overline{B}X - \underline{B}X$ is called the B-boundary region of X and it consists of those objects that cannot be classified with certainty as members of X with the knowledge in B. The set X is called "rough" (or "roughly definable") with respect to the knowledge in B if the boundary region is non-empty. Rough set theoretic classifiers usually apply the concept of rough set to reduce the number of attributes in a decision table [21] and to extract valid data from inconsistent decision tables. Rough set also accepts discretized (symbolic) input.

3. The computing process and empirical case study

Generally speaking, data mining is the process of analyzing data from different views and summarizing it into useful knowledge that can be used to predict revenue, costs, or both and much more. Technically, data mining is the process of finding correlations or patterns among dozens of attributes in large relational databases. Suitable domains for data mining are information-rich, a changing environment, no existing models, and knowledge-based decisions and can get high payoff for the right decisions. Then, due to the nature of the analytic problem, data preparation and cleaning of data pre-process are an often neglected but extremely important step in the data mining process. Major tasks in data preprocessing are data cleaning, data integration, data transformation, data discretization and data reduction, and so forth [4].

Referring to the work of [6], the process of data mining is fundamentally characterized as a three-stage iterative process: (1) The initial exploration, (2) Model building or pattern identification and validation, and (3) Deployment. Here we therefore present a research process (Figure 1.) to evaluate the good prediction tool for RGR of classification problems. The entire process of data-mining almost begins with the selection of data for the purpose of applying data-mining techniques, and ends with an analysis of the resulting information. That is, the task of extracting knowledge can be regarded as the iteration of these four steps. The proposed process is introduced as follows:

Step 1: Select the data
Step 2: Pre-process the dataset
Step 3: Analyze the dataset
Step 4: Evaluate accuracy and recommend

An empirical case study

A practical collected dataset is used in this empirical case study to demonstrate the proposed process: the dataset of RGR for 636 electronic firms in Taiwan stock trading system from 2004/03 to 2005/12 in Quarter. The dataset is split into two sub-datasets: one contains 2413 instances in 2004/03~2004/12; another contains 2490 instances in 2005/03~2005/12. Two sub-datasets are characterized by the following 13 attributes: (i) Total Fixed Assets, (ii) Gross Sales, (iii) Net Sales(Revenue), (iv) Cost of Goods Sold , (v) Gross Profit, (vi) Operating Expenses, (vii) Operating Income, (viii) Total Non-Operating Income, (ix) Total Non-Operating Expenses, (x) Total Salary, (xi) Total Depreciation, (xii) Numbers of Employee, and (xiii) Class; except class attribute, all attributes are continuous data in the dataset. The revenue growth rate is partitioned into three classes based on expert's experiment: Negative revenue growth rate (N), Positive revenue growth rate (P) and High Positive revenue growth rate (H_P). The Negative is defined as the firm which revenue growth rate is less than 0%, and Positive is defined as which revenue growth rate is from 0% to 100%, otherwise is classified as High Positive. The related computing process can be expressed as follows:

Step 1: Select the data

Select the data by extracting a portion of a large dataset from Taiwan stock trading system. Thus, a practical collected dataset is used in this empirical case study to demonstrate the proposed process: the dataset of RGR for 636 electronic firms in Taiwan stock trading system from 2004/03 to 2005/12 in Quarter. Moreover, the dataset is split into two sub-datasets: one contains 2413 instances in

2004/03~2004/12; another contains 2490 instances in 2005/03~2005/12.

Step 2: Pre-process the dataset.

Based on the dataset of RGR, we pre-process the dataset to suit for the proposed process. Major tasks in data preprocessing are data cleaning (e.g. fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies), data integration (combines data from multiple sources into a coherent store), data transformation (for example, normalization and aggregation) and data reduction (obtains reduced representation in volume but produces the same or similar analytical results).

Step 3: Analyze the dataset

In this study, the classification techniques in data mining for analyzing the dataset of RGR include Decision tree C4.5, Bayes net, Multilayer perceptron and Rough sets. Each technique has particular strengths, and is appropriate within specific situations in data mining depending on the properties of data. For example, Multilayer perceptron is very good at fitting highly complex nonlinear relationships. According to selected the important attributes in Step 2, analyze the dataset.

Step 4: Evaluate accuracy and recommend

For verification, each sub-dataset is split into two groups: the 67% dataset is used as a training set, and the other 33% is used as a testing set. The experiment is repeated ten times with the 67% / 33% random split. Table 1 presents the accuracy rate with standard deviation, comparison of different methods applying to same data in RGR sub-dataset of 2004. Table 2 presents the accuracy rate with standard deviation, comparison of different methods applying to same data in RGR sub-dataset of 2005. Table 3 shows partial rules by Rough set LEM2 in RGR sub-dataset of 2004, and Table 4 shows partial rules by Rough set LEM2 in RGR sub-dataset of 2005. The empirical results of two RGR sub-datasets indicate that the Rough sets outperforms the listing methods because of its improving accuracy and understandable rules.

4. Conclusions

A new process is proposed to compare and evaluate the results of different techniques in data mining for classifying problems. This process bases on using revenues, assets, profit, income, cost, and other data as condition attributes to determine the potential for future growth of its revenue by four tools of data mining techniques, Decision tree C4.5, Bayes net, Multilayer perceptron and Rough sets. The empirical results of two RGR sub-datasets indicate that the Rough sets outperforms the listing methods because of it improves accuracy and understandable rules. Moreover, the results may be useful for stock investors and system development and further researches. Specifically, the Rough sets method surpasses the listing method for three reasons: (1) In general, extracting rules based on Rough set LEM2 are superior to traditional methods because they deduce rule sets directly from data with symbolic and numerical attributes. (2) The accuracy rate and understandable rules demonstrate that the Rough sets approach outperforms the listing methods. (3) Rough set classifiers usually apply the concept of Rough set to reduce the number of attributes in a decision table [21], and data discretization is used to find the cut points for attributes.

References

[1] Jeffrey G. Covin, Kimberly M. Green, and Dennis P. Slevin, "Strategic process effects on the entrepreneurial orientation–sales growth rate relationship," Entrepreneurship Theory and Practice, pp. 57–82, 2006.

[2] Benjamin Graham, Security analysis, ISBN: 0071448209, McGraw-Hill Publisher, 3rd Edition, December 10, 2004.

[3] D. Hand, H. Mannila, and P. Smyth, Principles of data mining, MIT Press, Cambridge, MA, 2001.

[4] Han, J., and Kamber, M., Data mining: Concepts and techniques, New York: Morgan-Kaufman, 2000.

[5] Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R., "Advances in knowledge discovery & data mining", 1996.

[6] Edelstein, H., A., Introduction to data mining and knowledge discovery, (3rd ed), Potomac, MD: Two Crows Corp., 1999.

[7] John Sneed, "Estimating earnings forecasting model using fundamental analysis: Controlling for differences across industries," American Business Review, pp.17-24, 1999.

[8] Vanstone B., Finnie G., and Tan C., "Applying fundamental analysis and neural networks in the Australian stockmarket", Proceedings of the International conference on Artificial Intelligence in Science and Technology, Hobart, Tasmania, pp. 21-25, November 2004.

[9] Dunham, M.H., Data mining: Introductory and advanced topics, Prentice Hall, Upper Saddle River, NJ, 2003.

[10] Han, J., and Kamber, M., Data mining: Concepts and techniques, Morgan Kaufmann Publishers, San Francisco, 2001.

[11] Quinlan, J.R., "Induction of decision trees", Machine Learning, No.1(1), pp. 81-106, 1986.

[12] Quinlan, J.R., C4.5: Programs for machine learning, Morgan Kaufmann Publishers, San Mateo, CA, 1993.
[13] Jensen F.V., Bayesian networks and decision graphs, Springer, 2001.
[14] Leray P. and Francois O., "BNT structure learning package: Documentation and experiments", Technical report, Laboratoire PSI, Universite et INSA de Rouen, 2004.
[15] Murphy K.P. "The Bayes net toolbox for Matlab", Computing Science and Statistics, pp. 331-350, 2001.
[16] Verzilli C.J., Whittaker J.C., Stallard N. and Chasman D., "A hierarchical Bayesian model for predicting the functional consequences of amino-acid polymorphisms", Applied Statistics, 54, pp. 191-206, 2005.
[17] Conan-Guez B. and Rossi F., "Multi-layer perceptrons for functional data analysis: A projection based approach", ICANN 2002, Madrid, Spain, pp. 667-672, 2002.
[18] Rossi F., Conan-Guez B. and Fleuret F., "Functional data analysis with multi layer perceptrons", IJCNN 2002 (part of WCCI) proceeding, Honolulu, Hawaii, pp. 2843-2848, 2002.
[19] Bishop C. M., "A fast procedure for re-training the multi-layer perceptron", International Journal of Neural Systems, 2(3), 1991.
[20] Bedworth, M.D. and Lowe, D., "Fault tolerance in multi-layer perceptrons: a preliminary study", RSRE: Pattern Processing and Machine Intelligence Division, 1988.
[21] Pawlak, Z., "Rough sets, theoretical aspects of reasoning about data," Kluwer, Dordrecht, The Netherlands, 1991.
[22] Pawlak Z., "Rough sets," Informational Journal of Computer and Information Sciences, 11(5), pp. 341-356, 1982.
[23] K.P. Murphy, "Bayes net ToolBox," Technical Report, MIT Artificial Intelligence Labortary, 2002b. http://www.ai.mit.edu/~murphyk/.
[24] R. P. Lippmann, "An intorduction to computing with neural nets", IEEE ASSP Mag., pp. 4-22, April 1987.
[25] Bazan, J., Szczuka, M., RSES and RSESlib, "A collection of tools for rough set," Lecture Notes in Computer Science, Springer-Verlag, Berlin, pp. 106-113, 2005.

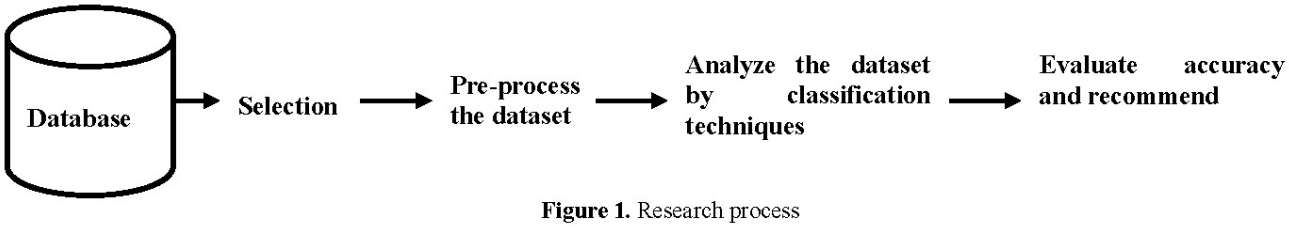

Figure 1. Research process

Table 1: Experiment results of RGR sub-dataset in 2004

Method	Testing Accuracy
Decision Tree-C4.5 [12]	71.64%
Bayes Net [23]	72.14%
Multilayer Perceptron [24]	72.14%
Rough set [25]	$75.15 \pm 1.6\%$

Table 2: Experiment results of RGR sub-dataset in 2005

Method	Testing Accuracy
Decision Tree-C4.5 [12]	63.13%
Bayes Net [23	62.53%
Multilayer Perceptron [24]	53.89%
Rough set [25]	$68.75 \pm 2.7\%$

Table 3: RGR sub-dataset result rule set example using Rough set LEM2 in 2004

Rules	Support
(Gross_Profit="(-lnf,7856899.84)")&(Cost_of_Goods_Sold="(265328.0,1196120.0)")&(Total_Non_Operating_Income="(6097.5,23429.0)")&(Total_Salary="(-lnf,28880.5)")&(Numbers_of_Employee="(-lnf,116.5)")&(Total_Depreciation="(-lnf,4225.0)") =>(Class=P)	19
(Gross_Profit="(-lnf,7856899.84)")&(Gross_Sales="(1292220.0,lnf)")&(Operating_Expenses="(159701.0,607656.0)")&(Cost_of_Goods_Sold="(1196120.0,1856529.92)")&(Numbers_of_Employee="(568.0,1542.0)") =>(Class=P)	17
(Gross_Profit="(-lnf,7856899.84)")&(Total_Salary="(28880.5,146108.0)")&(Operating_Expenses="(47228.0,159701.0)")&(Cost_of_Goods_Sold="(265328.0,1196120.0)")&(Total_Depreciation="(21526.5,130295.0)")&(Total_Non_Operating_Expenses="(9180.0,32323.5)")&(Numbers_of_Employee="(116.5,236.5)")=>(Class=P)	16
(Gross_Profit="(-lnf,7856899.84)")&(Total_Salary="(28880.5,146108.0)")&(Operating_Expenses="(47228.0,159701.0)")&(Cost_of_Goods_Sold="(265328.0,1196120.0)")&(Total_Depreciation="(21526.5,130295.0)")&(Numbers_of_Employee="(116.5,236.5)")&(Total_Non_Operating_Expenses="(9180.0,32323.5)")&(Gross_Sales="(752648.96,1292220.0)")&(Operating_Income="(56958.0,153635.0)")=>(Class=P)	13
(Gross_Profit="(-lnf,7856899.84)")&(Total_Salary="(28880.5,146108.0)")&(Cost_of_Goods_Sold="(265328.0,1196120.0)")&(Total_Depreciation="(21526.5,130295.0)")&(Total_Fixed_Assets="(739280.0,3764190.08)")&(Total_Non_Operating_Expenses="(9180.0,32323.5)")&(Operating_Income="(11810.5,56958.0)")=>(Class=P)	13

Table 4: RGR sub-dataset result rule set example using Rough set LEM2 in 2005

Rules	Support
(Operating_Expenses="(187456.0,1.680909952E7)")&(Gross_Profit="(650524.0,lnf)")&(Total_Salary="(342363.0,5198950.08)")&(Numbers_of_Employee="(714.5,3683.5)")&(Operating_Income="(661687.0,Inf)")&(Total_Fixed_Assets="(1113230.0,7571539.84)")&(Total_Depreciation="(168391.0,493096.0)") =>(Class=P)	15
(Gross_Profit="(51275.0,650524.0)")&(Total_Salary="(50094.0,342363.0)")&(Total_Depreciation="(11451.0,168391.0)")&(Net_Sales="(1476750.08,Inf)")&(Operating_Expenses="(187456.0,1.680909952E7)")&(Cost_of_Goods_Sold="(2101760.0,7625400.32)")&(Total_Non_Operating_Income="(98150.5,225313.0)")&(Total_Non_Operating_Expenses="(60052.0,127549.0)")=>(Class=P)	12
(Gross_Profit="(51275.0,650524.0)")&(Total_Salary="(50094.0,342363.0)")&(Total_Depreciation="(11451.0,168391.0)")&(Cost_of_Goods_Sold="(590418.0,2101760.0)")&(Net_Sales="(1476750.08,Inf)")&(Total_Fixed_Assets="(524796.0,1113230.0)")&(Operating_Expenses="(131530.0,187456.0)")&(Operating_Income="(177749.0,661687.0)")=>(Class=P)	9
(Net_Sales="(-Inf,1476750.08)")&(Gross_Profit="(51275.0,650524.0)")&(Total_Salary="(50094.0,342363.0)")&(Total_Depreciation="(11451.0,168391.0)")&(Cost_of_Goods_Sold="(590418.0,2101760.0)")&(Numbers_of_Employee="(118.5,356.0)")&(Operating_Income="(-Inf,2429.0)")&(Total_Fixed_Assets="(119152.0,309642.0)")=>(Class=H_P)	9
(Gross_Profit="(51275.0,650524.0)")&(Net_Sales="(-Inf,1476750.08)")&(Total_Salary="(12487.0,50094.0)")&(Numbers_of_Employee="(118.5,356.0)")&(Operating_Expenses="(35723.5,71076.0)")&(Total_Depreciation="(3159.0,11451.0)")&(Operating_Income="(40367.0,177749.0)")=>(Class=P)	8