Explorative Data Mining on Stock Data – Experimental Results and Findings

Lay-Ki Soon and Sang Ho Lee

Database Systems Laboratory, Department of Computing, School of Information Science, Soongsil University, 1-1 Sangdo-dong, Dongjak-gu Seoul, 156-743 Korea
laykisoon@gmail.com, shlee@comp.ssu.ac.kr

Abstract. This paper presents our preliminary explorative data mining experiments on stock data. The experiments are performed using information gain evaluation, chi-square test and decision tree, which aim to explore the underlying patterns of the stock dataset. The explorative patterns obtained show some unexpected outliers in deriving the influential stocks with regard to Kuala Lumpur Composite Index.

1 Introduction

Time-series data mining is identified as one of the 10 challenging problems in data mining research [11]. It can be focusing on trend analysis, similarity search, segmentation, clustering and classification [1, 2, 3, 9].

To gain better understanding on mining time-series data, we have embarked by performing preliminary data mining experiments on a stock dataset. The main objective of our experiments at this stage is explorative, which aims to understand the underlying patterns of the stock data. We have performed three experiments using information gain, chi-square test and decision tree. The first two experiments will generate a list of stocks which are influential to Kuala Lumpur Composite Index (KLCI), while the third experiment will produce classification rules, which denotes the inter-relationships among the stocks in terms of their trading performance with respect to KLCI.

This paper is organized as follows. Section 2 briefly discusses some related works which motivate our experiments. Section 3 presents the stock dataset and the data pre-processing. Section 4 to Section 6 explain the experiments, the results as well as our findings. Lastly, we conclude this paper in Section 7.

2 Related Works

There are numerous issues and challenges in mining time-series data to be addressed [1, 3, 7, 8]. Active research works have been done on the representation of time series data for classification purpose, such as Fourier transformation and weight vector. At this stage, our data mining experiments are purely descriptive by using the symbolic

R. Alhajj et al. (Eds.): ADMA 2007, LNAI 4632, pp. 562–569, 2007.
© Springer-Verlag Berlin Heidelberg 2007

representation of the stock dataset. In [9], sequential and non-sequential association rule mining (ARM) were used to perform intra and inter-stock pattern mining. Compared to [9], our exploratory mining experiments are similar to their inter-stock pattern mining, where associative relationships among different stocks were derived for inter-stock mining. In our work however, instead of deriving the expected association among the stocks, we are more interested in observing the performance of stocks with respect to a specific stock index, namely KLCI, which is always closely monitored in financial analyses.

3 Stock Dataset and Data Pre-processing

For the purpose of these experiments, we have used the stock trading dataset of Malaysia Exchange from December 31, 2001 till December 31, 2006, which amounts to 1233 days [4, 5]. Being one of the indices under Malaysia Exchange, KLCI is viewed as the local stock market barometer. It is a capitalization-weighted index, consisting 100 index components or stocks, where the list is updated every six months. KLCI's index components are selected based on several criteria, such as market capitalization, the company's performance, trading volume and so on [5].

The trading of the stock market within a day is recorded in a single text file. Figure 1 shows a snippet of a file. Each line represents the trading information of an index or a stock. Table 1 describes the information captured for the first line in Fig. 1.

```
COMPOSITE,D,12/29/06,1089.2800,1096.2400,1088.8000,1096.2400,9950,0
TRADING,D,12/29/06,151.3800,152.1300,151.1500,151.7400,1829,0
IND_PRODUCTS,D,12/29/06,91.5600,92.0700,91.3600,92.0700,886,0
PROPERTY,D,12/29/06,687.1400,693.4900,685.0600,693.4900,802,0
CONSTRTN,D,12/29/06,197.0300,200.0200,196.6100,200.0200,745,0
SECBOARD,D,12/29/06,91.5600,92.0300,91.3400,92.0300,713,0
FINANCE,D,12/29/06,8589.5100,8656.1200,8572.2400,8656.1200,363,0
PLANTATN,D,12/29/06,4215.6200,4315.9100,4215.6200,4315.9100,164,0
CONSUMER,D,12/29/06,266.5700,269.4600,265.5400,269.4600,140,0
MINING,D,12/29/06,526.8700,656.5700,526.8700,611.9900,124,0
INDUSTL,D,12/29/06,2206.8600,2225.4700,2204.4000,2225.4700,0,0
MESDAQ,D,12/29/06,118.8600,119.9200,118.4500,119.9200,238780100,0
TECHNOLOGY,D,12/29/06,28.3600,28.3600,27.9900,28.2800,7717700,0
7082WA,D,12/29/06,0.295,0.300,0.260,0.260,420,0
```

Fig. 1. Snippet of our stock dataset

Table 1. Information recorded at each line in the files

Information	Description
COMPOSITE	Kuala Lumpur Composite Index
D	Date
12/29/06	Dated December 29, 2006
1089.2800	Previous value (Last value of December 28, 2006).
1096.2400	Highest value of the day.
1088.8000	Lowest value of the day.
1096.2400	Closing / Last value of the day.
9950	Trading volume.
0	Constant.

Since noisy data is not an alien property of real-world datasets, we have spent almost 90% of our effort in these experiments on data pre-processing. Armed with the supporting information by Malaysia Exchange [5], we have developed a program to semi-automate the pre-processing activities.

The noise exists in our stock dataset can be generally classified into three main categories, which are duplicate records, inconsistent stock codes and incomplete files. Certain files contain duplicated records. For example, the transaction of stock 4863 is recorded twice in files dated December 9, 2002 and September 13, 2006. Thus, we have automated a thorough checking on all the files, ensuring only unique records on the same date.

In most of these files, stocks are represented by stock codes. For example, Telekom Malaysia Limited is recorded as 4863. However, in certain files, stocks are represented differently, such as TM or Telekom for Telekom Malaysia Limited. This inconsistency gets more complicated when the listed companies change their names or when companies merge. Having such inconsistencies, we have no choice but to rely on human intervention for the remedial action. Figure 2 depicts the steps in combating this problem. For the consistency of our experiments, we have chosen stock codes to represent the stocks, and COMPOSITE to represent KLCI. Note that *T* and *I* are provided by Malaysia Exchange for this purpose [5].

Input:
- Raw dataset *R* (1232 files with approximately 1300 records.)
- List of indices *I*, where each index *i* has its name $n_i \in N$.
- List of stocks *T*, where each stock *t* has its code $c_t \in C$ and name $s_t \in S$.

Output:
- List L containing valid stock codes and indices' names.
- Dataset D with stocks represented by stock codes and indices represented by names.

Automated Steps:

```
(1)   for every file in R
(2)         for every line in the file //refer Fig. 1
(3)               split every record according to commas
(4)               o = first item captured from step (3).
(5)               if (o = s_t, where s_t ∈ S)
(6)                     o = c_t   // c_t = stock code of s_t
(7)               end if
(8)               if (o ∈ C) or (o ∈ N)
(9)                     if (o ∉ L)
(10)                          L_i = o
(11)                          i++
(12)                    end if
(13)        end for
(14)  end for
```

Manual Checking:
- Given *T*, *I*, *L* and *D*, verify unique stocks and indices.

Fig. 2. Pre-processing steps for combating data inconsistency in the stock dataset

The files in the raw dataset contain all the market indices and stocks under Malaysia Exchange, where most of them contain approximately 1300 records. However, the file dated June 7, 2002 contains only stock indices without any record on stock trading. Up to the moment when this paper is prepared, the stock trading

information on this date is still not available. Therefore, we have disregarded the empty file in order to maintain the integrity of our mining results. As a result, instead of the initial 1233 files, we used only 1232 files for our experiments.

Let *st* denotes a stock code together with its performance and *C* represents all the stock codes in our dataset, every record in these files is converted into the following format:

$$st: c = v \tag{1}$$

where $c \in C$ and $v \in \{UP, DOWN, SAME\}$. *UP*, *DOWN* and *SAME* denote positive, negative and unchanged performance of a stock respectively. For example, the record shown in Table 1 is converted to COMPOSITE=UP. For every file in the dataset, after converting each record into format (1), we then combined them into one row, separated by commas. Every file was formatted as a single row composing the stock trading on that same day. These rows were then inserted into a file, which eventually contains 1232 rows of data, as illustrated in Fig. 3.

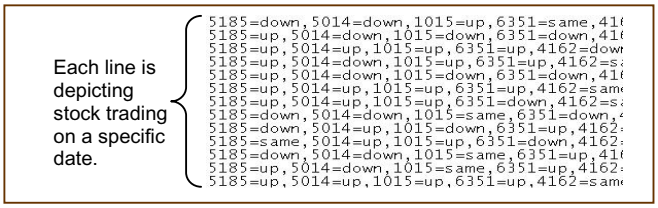

Fig. 3. Converted dataset, which is ready for the experiments

4 Information Gain Evaluation

Since KLCI serves as the benchmark for Malaysia equity market, we are particularly interested to identify influential stocks with respect to KLCI. Information gain is an attribute selection measure based on entropy, where it studies the value of certain piece of information [2]. This experiment uses information gain to derive a set of stocks which are sorted according to their collective worth with respect to KLCI within 31 December 2001 to 31 December 2006. The stocks with higher information gain imply greater worth or influence against KLCI. We used the *Select Attribute* feature provided by Weka [10], with 10-fold cross validation for this experiment.

Intuitively, the 100 stocks which constitute KLCI should appear as the top 100 stocks by information gain with respect to KLCI. However, some of the results generated by the experiment are not as expected. Out of the top 100 stocks by information gain, only 38 of them constitute the current KLCI. Table 2 shows the top 20 stocks by information gain with respect to KLCI. Since market capitalization is one of the criteria in selecting KLCI's constituents, it is listed for every company in the third column. For example, AMMB is the 23rd highest companies in terms of market capitalization in Malaysia Exchange, which is MYR6775.2 million [5]. TIME is not one of the top 100 companies by market capitalization, but it is included in current KLCI and has the market capitalization of MYR519.41 million.

Although some of these stocks are not KLCI's constituents, by the statistical analysis of information gain, their performance throughout this period of time are considered as worthy to KLCI's performance. Several reasons may lead to the unexpected results. One is because the dataset spread across 31 December 2001 to 31 December 2006. Hence, the experimental results represent the order of stocks by their collective information gain with respect to KLCI within this period of time.

Out of the top 20 stocks (Table 2), only 4 of them do not constitute the current KLCI. We consider LEADER and MRCB as exceptional cases as they belong neither to KLCI nor the top 100 companies by market capitalization [5]. Nevertheless, both of these companies have been performing steadily. Market analyses considered the performance of LEADER, which has MYR257.5 million of market capitalization is ahead of expectation. Although not included in KLCI, AMMB-WA and PBBANK-FOREI are sister companies of AMMB and PBBANK respectively.

Table 2. Top 20 Stocks by Information Gain Evaluation

No.	Stock	Status	Remarks
1	AMMB	KLCI	Market Cap.: 23rd
2	RHBCAP	KLCI	Market Cap.: 24th
3	TM	KLCI	Market Cap.: 3rd
4	MAYBANK	KLCI	Market Cap.: 2nd
5	GENTING	KLCI	Market Cap.: 8th
6	COMMERZ	KLCI	Market Cap.: 5th
7	TENAGA	KLCI	Market Cap.: 1st
8	LEADER	Non-KLCI	Exceptional
9	MRCB	Non-KLCI	Exceptional
10	PBBANK	KLCI	Market Cap.: 6th
11	AMMB-WA	Non-KLCI	Sister company of AMMB
12	YTL	KLCI	Market Cap.: 17th
13	MULPHA	KLCI	Market Cap.: 81st
14	MAS	KLCI	Market Cap.: 28th
15	BJTOTO	KLCI	Market Cap.: 24th
16	PBBANK-FOREI	Non-KLCI	Sister company of PBBANK
17	PLUS	KLCI	Market Cap.: 13th
18	DRBHCOM	KLCI	Market Cap.: 82nd
19	MAGNUM	KLCI	Market Cap.: 46th
20	TIME	KLCI	Market Cap.: MYR519.41 Mil.

In short, this experiment has sorted the stocks according to their worth or influence with regard to KLCI. We are surprised to observe that market capitalization of the stocks and their influence capabilities to KLCI may not be directly correlated.

5 Chi-Square Test

In this second experiment, we have applied χ^2 (chi-square) test to discover the statistical correlation between every stock and KLCI. However, note that correlation analysis does not imply causality [2]. Table 3 shows the contingency table between stock AMMB and KLCI. The columns represent the distinct values of AMMB. NULL

value is possible when the trading information of AMMB is not available in certain trading days. The first and third rows represent the distinct values of KLCI. Out of 1232 days of stock trading, there are 372 days where both AMMB and KLCI performed positively; and 368 days where both of them performed negatively. The chi-square test for the correlation relationship between AMMB and KLCI can be computed as

$$\chi^2 = \frac{(372-254.16)^2}{254.16} + \frac{(171-284.81)^2}{284.81} + ... + \frac{(0-0.47)^2}{0.47} = 213.56$$

Table 3. Contingency Table Between KLCI and AMMB

KLCI \ AMMB	UP	DOWN	SAME	NULL	Total
UP	372	171	107	1	651
UP-Expected Frequency	*254.16*	*284.81*	*111.49*	*0.53*	
DOWN	109	368	104	0	581
DOWN-Expected Frequency	*226.84*	*254.19*	*99.51*	*0.47*	
Total	481	539	211	1	1232

By convention, we can reject the hypothesis that AMMB and KLCI are independent if its χ^2 value exceed a threshold (significance value). However, as we intend to explore the chi-square value of all the stocks, we do not apply any threshold in this experiment. We have used 10-fold cross validation in this experiment [10].

Table 4. Top 20 Stocks By Chi-Square Test

No.	Stock	Status	Remarks
1	AMMB	KLCI	Market Cap.: 23rd
2	RHBCAP	KLCI	Market Cap.: 24th
3	TM	KLCI	Market Cap.: 3rd
4	MAYBANK	KLCI	Market Cap.: 2nd
5	GENTING	KLCI	Market Cap.: 8th
6	COMMERZ	KLCI	Market Cap.: 5th
7	TENAGA	KLCI	Market Cap.: 1st
8	*MRCB*	*Non-KLCI*	*Exceptional*
9	*LEADER*	*Non-KLCI*	*Exceptional*
10	PBBANK	KLCI	Market Cap.: 6th
11	AMMB-WA	Non-KLCI	Sister company of AMMB
12	YTL	KLCI	Market Cap.: 17th
13	MULPHA	KLCI	Market Cap.: 81st
14	MAS	KLCI	Market Cap.: 28th
15	BJTOTO	KLCI	Market Cap.: 24th
16	PBBANK-FOREI	Non-KLCI	Sister company of PBBANK
17	*DRBHCOM*	*KLCI*	*Market Cap.: 82nd*
18	*PLUS*	*KLCI*	*Market Cap.: 13th*
19	*TIME*	*KLCI*	*Market Cap.: MYR519.41 Mil.*
20	*UEMBLDR*	*Non-KLCI*	*Market Cap.: 89th*

Table 4 shows the top 20 stocks sorted according to their chi-square values. Comparing the results from information gain and chi-square test, different stocks are presented at 8^{th}, 9^{th}, and 17^{th} – 20^{th}. Likewise, 8^{th} and 9^{th} stocks are considered as

anomalies which may trigger further financial analysis. One significant difference between the lists generated by information gain and chi-square is on MAGNUM (19th in Table 2) and UEMBLDR (20th in Table 4). MAGNUM is placed at 22nd by chi-square test. Despite not being as one of KLCI's constituents, UEMBLDR is the 89th highest company in Malaysia Exchange by market capitalization.

Nevertheless, the overall results generated by both information gain and chi-square test are generally consistent, which can be considered as a positive verification of the results from both experiments. Again, we can conclude that market capitalization and the performance of KLCI are not directly correlated in most of the cases.

6 Decision Tree

Different from the previous two experiments, we only select the 100 index components of current KLCI in this experiment. Having KLCI as the class attribute, we expect to derive a set of classification rules. Our experiment on DT is performed using the implementation of C4.5 in Weka [6, 10]. The best accuracy – 74.51% is obtained at 14-fold cross validation.

Table 5. Classification Rules Derived from Decision Tree

No.	Classification Rules
1	*if* AMMB=UP *and* RHBCAP=UP *and* TIME=UP *then* COMPOSITE=UP **(191, 8)**
2	*if* AMMB=DOWN *and* MBB=DOWN *and* MAS=DOWN *then* COMPOSITE=DOWN **(118, 5)**
3	*if* AMMB=DOWN *and* MBB=SAME *and* COMMERZ=DOWN *then* COMPOSITE=DOWN **(80, 12)**
4	*if* AMMB=DOWN *and* MBB=UP *and* MMBCORP=DOWN *and* TA=DOWN *then* COMPOSITE=DOWN **(38,4)**
5	*if* AMMB=UP *and* RHBCAP=UP *and* TIME=SAME *then* COMPOSITE=UP **(36, 3)**
6	*if* AMMB=SAME *and* MAS=UP *and* YTL=UP *then* COMPOSITE=UP **(34, 1)**
7	*if* AMMB=UP *and* RHBCAP=DOWN *and* TENAGA=UP *and* GENTING=UP *then* COMPOSITE=UP **(28)**
8	*if* AMMB=UP *and* RHBCAP=UP *and* TIME=DOWN *and* BERNAS=UP *then* COMPOSITE=UP **(22)**
9	*if* AMMB=SAME *and* MAS=DOWN *and* GUTHRIE=DOWN *then* COMPOSITE=DOWN **(18)**
10	*if* AMMB=UP *and* RHBCAP=DOWN *and* TENAGA=DOWN *and* DRBHCOM=DOWN *and* COMMERZ=DOWN *then* COMPOSITE=DOWN **(18)**

With 220 leave nodes in the decision tree, it composes 293 classification rules. Based on the number of records covered by these rules, the top 10 rules are listed in Table 5. The figures followed after each classification rule indicate the number of records correctly and incorrectly classified by the rule respectively. Having the highest information gain, AMMB is selected as the root node. In the second level, the selected stocks are RHBCAP, MBB and MAS. All these 4 stocks are blue chip stocks which have high market capitalization. Using these classifications rules, we may observe the inter-relationships among these KLCI constituents as well.

7 Conclusion and Future Works

In this paper, we have presented our experience and findings on exploratory stock data mining using information gain, chi-square test and decision tree. Extensive data pre-processing steps has been performed. Within the generated underlying patterns, we obtained some unexpected results, which trigger further investigation by the financial analysts. With proper data pre-processing, we believe that these insightful patterns will be of valuable knowledge for stock market analysis.

For our future works, we plan to explore the temporal semantics of the dataset, such as the evolution of the correlation relationships among the stocks over a period of time. Having the patterns of stock data captured over time, we may perform similarity search among the stocks. Besides, the possibility of conceptualizing the relationships among stocks and indices ontologically will be investigated.

Acknowledgments. This work was supported by Seoul R&BD Program (10581cooperateOrg93112). We also thank Mr. Alex Tham Poh Seng and Mr. Abdul Rahim Hussein from Malaysia Exchange for providing us the information and advice throughout these experiments.

References

1. Antunes, C.M., Oliveira, A.L.: Temporal Data Mining: An Overview. In: Proceedings of KDD Workshop on Temporal Data Mining, San Francisco, pp. 1–13 (2001)
2. Han, J., Kamber, M.: Data Mining Concepts and Techniques. Morgan Kauffman Publishers, Elservier, San Francisco, CA (2006)
3. Keogh, E., Kasetty, S.: On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration. In: International Journal of Data Mining and Knowledge Discovery, vol. 7(4), pp. 349–371. Springer, Netherlands (October 2003)
4. Malaysia Exchange Dataset. (subscription) Available: http://www.klsedaily.com
5. Official Website of Malaysia Exchange. Available: http://www.bursamalaysia.com
6. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kauffman, San Francisco (1993)
7. Roddick, J.F., Hornsby, K., Spiliopoulou, M.: An Updated Bibliography of Temporal, Spatial, and Spatio-Temporal Data Mining Research. In: Roddick, J.F., Hornsby, K. (eds.) TSDM 2000. LNCS (LNAI), vol. 2007, pp. 147–164. Springer, Heidelberg (2001)
8. Roddick, J.F., Spiliopoulou, M.: A Survey of Temporal Knowledge Discovery of Paradigms and Methods. IEEE Transactions on Knowledge and Data Engineering 14(41), 750–767 (2002)
9. Ting, J., Fu, T., Chung, F.: Mining of Stock Data: Intra- and Inter-Stock Pattern Associative Classification. In: Proceedings of 2006 International Conference on Data Mining, Las Vegas, USA, June 2006, pp. 30–36 (2006)
10. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)
11. Yang, Q., Wu, X.: 10 Challenging Problems in Data Mining Research. International Journal of Information Technology & Decision Making 5(4), 1–8 (2006)