

Mining of Stock Data: Intra- and Inter-Stock Pattern Associative Classification

Jo Ting, Tak-chung Fu, and Fu-lai Chung[†]

Department of Computing, Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong.

Abstract—In this paper, a pattern-based stock data mining approach which transforms the numeric stock data to symbolic sequences, carries out sequential and non-sequential association analysis and uses the mined rules in classifying/predicting the further price movements is proposed. Two formulations of the problem are considered. They are intra-stock mining which focuses on finding frequently appearing patterns for the stock time series itself and inter-stock mining which discovers the strong inter-relationship among several stocks. Three different methods are proposed for carrying out associative classification/prediction, namely, Best Confidence, Maximum Window Size and Majority Voting. They select the mined rule(s) and make the final prediction. A modified Apriori algorithm is also proposed to mine the frequent symbolic sequences in intra-stock mining and the frequent symbol-sets in inter-stock mining. Various experimental results are reported.

Keywords: Stock Data Mining, Time Series Analysis, Association Rule Mining, Associative Classification

I. INTRODUCTION

Armed with better information, the management of a company can apply their creativity and judgment to make better decisions and get better returns. Knowledge discovery in databases (KDD) is now well-known for its potential in nontrivial extraction of implicit, previously unknown and potentially useful information in data [1]. It has attracted tremendous interest in the research community as well as commercial market place and has emerged as an automated mechanism for better understanding and characterization of data. The KDD process generally involves a few processing steps, namely, data selection, feature-value selection [2] and transformation [3], data mining, presentation and evaluation. As data mining is the core component of the KDD process, the term data mining and KDD have been used interchangeably by many researchers [4,5].

Mining on financial data is not trivial. In the financial domain, technical analysis [6] is one of the most commonly used methods for predicting price movements and future market trends. By studying charts of past market actions which take into account of prices of instruments and volume of trading, one may obtain certain suggestions on investment decisions [7]. Technical analysis concerns with what has

actually happened in the market, rather than what should happen. It is subjective and its success highly depends on the analyst's experience. As an objective tool, data mining can be applied to discover the interesting behavior within a time series or the relationship among a set of time series so that investors can collect more useful information from the already available but huge amount of data. For example, looking for repetitive patterns in a stock time series can be very useful for stock investors.

In this paper, a pattern-based stock data mining problem is considered. The patterns being referred to are the typical technical (analysis) patterns like Yin (bearish) and Yang (bullish). We study how the sequential and non-sequential association rule mining [8,9] are applied to stock time series applications. Two different formulations of the problem are targeted, namely, intra-stock pattern mining and inter-stock pattern mining and they are elaborated as follows.

Intra-stock pattern mining concerns with the discovery of repetitive temporal association patterns for the stock itself across a time span, e.g., 5 trading days. As shown in Fig.1, three patterns have been boxed and they can be interpreted as having a few continuous bearish (going down) days and then followed by a few bullish (going up) days. By converting the numerical stock time series data into symbolic sequence based on the concept of technical or candlestick analysis, the sequential association rules mined can characterize such kind of market observations for a selected stock.

While intra-stock pattern mining is to find the sequential association rules within a time series, inter-stock pattern mining picks several stocks and finds the relationship (association) among them. Inter-stock mining can be used to find the non-sequential association of stock symbols/patterns within a trading interval (e.g. the same trading day). With such formulation, the inter-relationship of stocks from different sectors, e.g., an oil supply stock 0883* and an airline stock 0293 as shown in Fig.2 can be studied. It is very sensible that when oil price increases, the operational cost of airline services increases, thus having negative influence on the stock price of the airline stock. Moreover, the inter-relationship between mother/sister companies is an interesting area to explore. As exemplified in Fig.3 for stock 0001 and stock 0013, their price movements are pretty similar, with up trends and down trends in the same pace.

[†] Corresponding author: cskchung@comp.polyu.edu.hk

* stock code used by the Hong Kong Stock Exchange



Figure 1 - Intra-Stock Pattern Mining for stock 0005 (Bank)

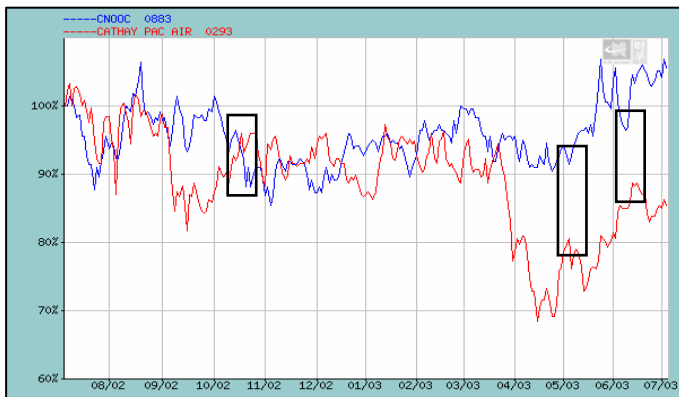


Figure 2 - Inter-Stock Pattern Mining for stock 0883 (Oil) and 0293 (Airline)

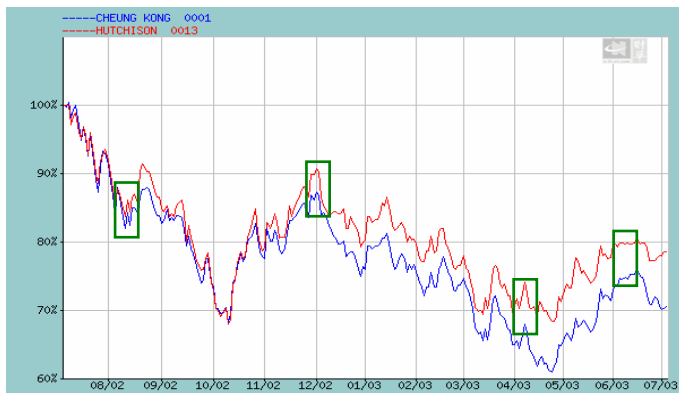


Figure 3 - Inter-Stock Pattern Mining for stock 0001 (Mother) and 0013 (Sister)

In order to apply the intra- and inter-stock pattern mining results, we consider the problem of how to use the sequential and non-sequential association rules in prediction, i.e., associative classification [10]. In the next section, we first describe the data preprocessing step necessary for association rule mining of numeric stock data and then introduce how association rules, including both sequential and non-sequential ones, are used in stock prediction/classification. Sections III and IV describe the intra-stock and inter-stock pattern mining processes respectively. In Section V, the experimental results are reported.

II. DATA PRE-PROCESSING AND ASSOCIATIVE CLASSIFICATION

A. Data Pre-processing

Time series data are difficult to manipulate, but when they can be treated as symbols (item units) instead of data points, interesting patterns can be discovered and it becomes an easier task to mine them. Thus, it is suggested to convert the basic unit into symbols, i.e., numeric-to-symbolic conversion. The idea in technical analysis is borrowed and is flexible enough to discover knowledge of varying precision levels and comprehensibility, depending on the user's problem-specific goals. The numeric-to-symbolic conversion transforms the available features (e.g. Open, High, Low, Close prices) of a financial instrument into a string of symbols. In other words, the numeric data sequences from each stock time series are interpreted and a unique symbol is then used to label them individually. Such a conversion process can be extended to granulate the numerical data into different time granularities and it provides a large collection of symbol strings, hopefully at various time granularities, which can then be used for different applications.

Since our focus is on stock market, we make use of the *Open*, *High*, *Low* and *Close* prices to carry out the numeric-to-symbolic conversion. Firstly, the resolution of the basic unit must be defined. The most common unit is in daily scale, while it is easy to extend the above price measurements to other resolutions, say, weekly, monthly, etc. Here, one of the challenges being faced is to determine an appropriate number of symbols that is representative and also flexible enough for different time series. If the number of symbols is too many, then the occurrence of each symbol would be infrequent, making the mining process and the subsequent prediction task difficult. Even the rule can be generated with high confidence, say 100%, the pattern may not happen again and hence the rule is useless. On the other hand, if the number of symbols is too few, the support of each symbol would increase but the confidence may not be high enough and the interestingness of the mined rules is questionable.

After many experimental proofs, features more than 3 symbols would yield very little occurrence for the rules and those interesting rules never happen again. So, in this paper, only 1 feature is taken, i.e., the price movement consisting of 3 values/possibilities:

Symbol	Definition
Yang	$(Close - Open)/Close > Threshold$
Yin	$(Open - Close)/Close > Threshold$
Level	$ Close - Open /Close \leq Threshold$

E.g., for $Open=\$100$, $Close=\$99.5$ and $Threshold=1\%$, a level feature value will be generated because

$$\frac{|99.5 - 100|}{99.5} < 1\% .$$

The *Threshold* is a user-defined parameter. If it is set to 0%, then the symbol or feature value for the example above would be "Yin" instead.

B. Associative Classification

In our experiments, 6 years of daily stock time series data were used and 70% of them were for *training* while the remaining 30% were for *testing*. After the association rules are generated from the training data, they are used to classify the stock price movements. The problem is not straightforward and some issues need to be resolved. For example,

- Which rule or set of rules should be selected for making prediction?
- If the consequents of the selected rules are different from each other, which one(s) should be adopted?
- How to combine the predictions from multiple rules and generate the final prediction?

In this paper, three different methods are proposed, namely, *Best confidence*, *Maximum Window Size* and *Majority Voting*. They are elaborated as follows.

▪ *Best Confidence:*

In the best confidence approach, the rule with the highest confidence among all the mined rules matching the fact of the testing data is selected for classifying the testing data. Generally speaking, higher confidence should yield better prediction. For example in sequential association analysis, a testing data sequence “Level→Level→Level→Level”, corresponding to a window size equal to 4, the rule R3 among the four matched rules in Table I below will be selected because its confidence is the highest one and the classification will be “Yang”.

Table I. An example for the best confidence approach

Rule	Window Size	Mined Rule (Antecedent -> Consequent)	Confidence %
R1	1	Level -> Level	34
R2	2	Level-Level -> Yin	60
R3	3	Level-Level-Level -> Yang	70
R4	4	Level-Level-Level-Level -> Yin	65

▪ *Maximum Window Size:*

The basic idea here is to choose the mined rule with the longest antecedent matched with the testing data sequence. Generally speaking, the support of the rules with smaller window size is higher than that of the rules with larger window size. From the example in Table II, when the window size increases, the combination of symbols increases exponentially and thus the support is generally lower. For such an example, though R3 has the highest confidence, R120 is chosen for prediction in the maximum window size approach.

▪ *Majority Voting:*

The rationale behind the majority approach is to get synergy from the set of mined rules matched with the testing data sequence. Majority voting is perhaps the most typical choice and is adopted here. With the example shown in Table III, the mined rules, with any window size, which match with the antecedent of the testing data sequence, i.e., “Yin-Yin-Yin”, would be selected. There are eight rules here and according to our majority approach the classification result should be “Level”. It is because majority vote for “Yin” is $20 + 15 + 40 = 75$, majority vote for “Yang” is $30 + 25 = 55$, whereas majority

vote for “Level” is $50 + 60 + 60 = 170$.

Table II. An example for the maximum window size approach

Combination	Window Size	Rule	Mined Rule – Antecedent Portion	Confidence %		
$3^1 = 3$	1	R1	Yin	33		
		R2	Yang	33		
		R3	Level	34		
$3^2 = 9$	2	R4	Yin-Yin	12		
		R5	Yin-Yang	13		
		R6	Yin-Level	12		
		R7	Yang-Yin	10		
		R8	Yang-Yang	10		
		R9	Yang-Level	11		
		R10	Level-Yin	10		
		R11	Level-Yang	10		
		R12	Level-Level	12		
		$3^3 = 27$	3	R13	Yin-Yin-Yin	4
				R14	Yin-Yin-Yang	3
				R15	Yin-Yin-Level	4
...	...					
R39	Level-Level-Level			20		
$3^4 = 81$	4	R40	Yin-Yin-Yin-Yin	2		
				
		R120	Level-Level-Level-Level	17		

Table III. An example for the majority approach

Testing Rule	Item Size	Antecedent	Consequence	Confidence	Count
R1	1	Yin	Yin	0.4	20
R2		Yin	Yang	0.2	30
R3		Yin	Level	0.4	50
R4	2	Yin-Yin	Yin	0.5	15
R5		Yin-Yin	Yang	0.7	25
R6		Yin-Yin	Level	0.2	60
R7	3	Yin-Yin-Yin	Yin	0.5	40
R8		Yin-Yin-Yin	Level	0.6	60

III. INTRA-STOCK PATTERN MINING

For intra-stock data mining, we are looking for repetitive pattern on the selected stock itself and generate association rules based on the symbolic pattern with user specified minimum support and confidence. For the symbol sequence obtained by the numeric-to-symbolic process

XXXXXABCXXXABCXXABCXXXXXXABC

where A, B and C denote the symbols/signs of interest, e.g. Yin, Level, and Yang respectively, the sequence ABC occurred 4 times. If it is found to be frequent, one may conclude that the stock price going down and then level implies a price rebound in the following time period (e.g. trading day).

Unlike the traditional data mining process where the total number of transactions is fixed, the support count for our intra-stock data mining is calculated based on a sliding window concept. For example, for 1000 consecutive stock prices (say from 1000 trading days), we have

$$n\text{-symbol itemset} = 1000 - n + 1$$

Therefore, in the first pass of the algorithm, it simply counts the number of occurrences of each item (symbol) to determine the frequent 1-itemsets. The subsequent passes consist of two phases.

- i) First, the frequent itemsets (frequently appearing symbol patterns) L_{k-1} found in the $(k-1)^{th}$ pass are used to generate the candidate itemsets C_k , using the apriori-gen function [8].
- ii) Second, the symbol string S is scanned using sliding window with length k and the support of each candidate itemset in C_k is calculated as following:

$$support(C_k) = \frac{c.count}{length(S) - k + 1}$$

where $c.count$ denotes the number of occurrences of C_k .

A. Modified Apriori Algorithm

In this section, the modifications being made on the Apriori algorithm [8] to suit our mining task are described. The traditional algorithm considers a set of transactions, where each transaction is a set of un-ordered items. However, symbol strings are our focus and frequent itemsets (frequently appearing symbolic patterns) are to be discovered. The traditional Apriori algorithm does not consider the temporal order of the items, i.e., ABC and CBA are treated as the same itemset while in our mining task, they should be referred to different patterns. The sequence of items appeared in a pattern will be considered in our modified algorithm.

In addition, the traditional Apriori algorithm does not consider the self-joining of item itself when forming a candidate itemset, i.e., item A would not be self-joined with itself to form AA. However, in time series data mining, consecutive appearance of the same symbol may happen, e.g., 3 consecutive bullish days would lead to a bearish day. The item ‘‘bullish day’’ repeats for 3 times in such a pattern. In order to illustrate our modified Apriori Algorithm, an example is referred. It consists of 100 consecutive days of stock prices and the total number of symbols is equal to 3, e.g. Yin (A), Level (B) and Yang (C).

Frequent 1-Itemset

Just to pick those symbols satisfying the minimum support. Assume that the frequent 1-itemsets are A, B, C.

Candidate 2-itemset

Then, the candidate 2-itemsets are generated as the following table. By using our modified algorithm, more candidates will be generated.

Table IV. Joining Symbols (2-itemset) in Modified Apriori Algorithm

How to Join Symbols	Result
Traditional Join	AB, AC, BC
Inverted Join - Due to the importance of temporal order in time series rule mining	BA, CA, CB
Self Join - Due to the possibility of repeated occurrence of the same symbols in a pattern	AA, BB, CC

Frequent 2-itemset

As mentioned, with the concept of sliding window, the

support count should be computed appropriately. We pick those candidate 2-itemsets satisfying the minimum support as the frequent 2-itemsets.

Candidate 3-itemset

Assume the frequent 2-itemsets are AA, AB, BA, and BB. For the traditional Apriori algorithm, the candidate 3-itemsets should be generated by checking the left-most symbol of the frequent 2-itemsets. If they are the same, then join the corresponding frequent 2-itemsets together, e.g., $\overline{A}A$ and $\overline{A}B$ would become $\overline{A}AB$. However, when we carry out classification/prediction based on the association rules, the most recent symbols (historical data) should be used. Thus, we would rather detect the right-most symbol, in our example, $A\overline{A}$ and $B\overline{A}$ would generate $AB\overline{A}$. Inverted join starting from 3-itemset can be eliminated, due to the fact that once the inverted symbols are generated in 2-items, they would be brought into later iteration if small min support is used. Thus, we can form candidate 3-itemset as follows:

Table V. Joining Symbols (3-itemset) in Modified Apriori Algorithm

How to Join Symbols	Result
Right-Most Join	ABA, ABB
Self Join	AAA, AAB, BBA, BBB

Frequent 3-itemset

By using the concept of sliding window, we pick those candidate 3-itemsets which satisfied the min support.

Candidate 4-itemset

Assume the frequent 3-itemsets are ABA, BBA. Similar to the candidate 3-itemset generation step, here we check the 2 right-most symbols. If matched, we join the frequent 3-itemsets together to form candidate 4-itemset by:

Table VI. Joining Symbols (4-itemset) in Modified Apriori Algorithm

How to Join Symbols	Result
Right-Most Join	ABBA
Self Join	AABA, BBBA

The iteration above repeats until no symbols can be joined.

IV. INTER-STOCK PATTERN MINING

While intra-stock pattern mining is to find the association rules within a time series, inter-stock rule mining concerns with several stocks among which the patterns (associations) are mined. Since several stocks can be picked, we can find the inter-relationships of stocks from:

- 1) same industrial domain, e.g. banking, telecommunication
- 2) manufacturing-chain industrial domain, e.g. coal and town gas, cotton, fashion
- 3) mother and sister company

A. The Mining Process

For inter-stock data mining, basically the association of different stocks’ price symbols for a selected time granularity, say a trading day, is studied. Thus, each trading day can be considered as a transaction and the involved price symbols can be treated as items, under the association analysis framework.

Let's take a 3-stock association analysis as an example. After the numeric-to-symbolic conversion, we can form transactions like Table VII below.

Table VII. Numeric-to-symbolic Converted Stock Transactions

Trading Day	Stock 0001	Stock 0013	Stock 0008
2005/11/1	A	B	A
2005/11/2	A	A	B
2005/11/3	A	A	A
...			
2005/11/28	B	A	A
2005/11/29	A	A	A
2005/11/30	C	C	A

Since the sliding window concept does not exist, the number of transactions is equal to the number of trading days selected. The idea can be easily extended to trading weeks, trading month and multiple stocks. It is a typical non-sequential association analysis problem and no temporal information has to be incorporated.

In the following, a 3-stock association analysis example is used to illustrate the mining process. There are 100 transactions and 3 symbols (Yin, Yang and Level) are used.

Frequent 1-Itemset

Just to pick those symbols satisfied the minimum support. Assume the frequent 1-itemset are A, B, C, i.e. all the 3 symbols are frequent.

Candidate 2-itemset

The total number of candidate 2-itemset is N^2 , where N is the number of frequent 1-itemset. Hence, we have 9 candidate 2-itemsets: AA, BB, CC, AB, AC, BC, BA, CA, CB. These symbol-sets or itemsets are attached with: *Stock 1 & Stock 2*, or *Stock 1 & Stock 3*, or *Stock 2 & Stock 3*

Frequent 2-itemset

Then we pick those symbols which satisfied the min support

Candidate 3-itemset

The total number of candidate 3-itemset is N^3 , i.e. 27 in our example:

AAA, AAB, AAC, ABA, ABB, ABC, ACA, ACB, ACC, BAA, BAB, BAC, BBA, BBB, BBC, BCA, BCB, BCC, CAA, CAB, CAC, CBA, CBB, CBC, CCA, CCB, CCC

which are attached with the three stocks respectively.

Frequent 3-itemset

Similarly, we pick those 3-itemsets satisfying the minimum support.

Such kind of iteration continues until the frequent item size equals to the number of stocks you picked.

B. Predictive Inter-stock Mining

The inter-stock time series mining problem introduced previously focused on the relationship among stocks within the same transaction period, i.e. intra-transaction mining. It cannot be used to predict the future price movement, e.g., predicting the price symbol of Stock 3 tomorrow (bullish or bearish) given that Stock 1 is Level and Stock 2 is Yin today. Such an idea can be extended to obtain rules like:

- i) Candlestick of Stock 1 and Stock 2 is Yin, then Stock 3 tomorrow will be Yang

- ii) Candlestick of Stock 2 is Yin and Stock 3 is Yang, then Stock 1 tomorrow will be Yang
- iii) Candlestick of Stock 1 is Level and Stock 3 is Level, then Stock 2 tomorrow will be Level too

The difference between predictive inter-stock mining and non-predictive one is how to extract symbols from the stock time series data. In the experiments to be reported in the next section, we will focus on using today's stock price symbols to predict tomorrow's price movement.

V. EXPERIMENTAL RESULTS

Due to the limited paper length, only some of the experiment results are reported. All experiments have adopted the general process below for prediction or classification.

- i) Partition the full dataset into training one and testing one.
- ii) By using the training data set to perform numeric-to-symbolic conversion, apply the Apriori or modified Apriori Algorithm to generate the association rules.
- iii) Perform numeric-to-symbolic conversion on the testing data set.
- iv) Based on the mined association rules that match with the facts (of the testing data), apply the best confidence, maximum window size and majority methods to carry out associative classification.
- v) When the consequence (Right-Hand-Side symbol) predicted/classified by the selected mined rule(s) matches with the actual consequence from the testing data, the prediction or classification is defined as "Correct"; otherwise it is defined as "Wrong". Moreover, two special results are noted here.
 - When the confidence of the mined rule(s) (from the training dataset) is not sufficient to support the prediction or classification, the result is defined as "No Prediction" and this prediction or classification is not counted.
 - When the antecedents/facts/patterns from the testing data cannot be matched by the mined rule, the result is defined as "No Prediction" and again, this prediction or classification is not counted as well.

Thus, the prediction/classification accuracy is defined as:

$$Accuracy = \frac{No. of Wrong Prediction}{No. of Correct Prediction + No. of Wrong Prediction}$$

A. Performance of Associative Classification Methods in Intra-stock Mining

In this experiment, 15 stocks were selected to test on the three associative classification methods, i.e., Best Confidence, Maximum Window Size (Maximum Item Size) and Majority Voting, in intra-stock mining. Among the 15 stocks, the majority voting method yields the best classification result for 9 times and hence is better than the other two methods. Although a rule having the highest confidence is expected to provide better prediction, it may not be effective or representative if its support is not high enough. This problem is also shared by the maximum window size method because the support can be very low when the window size is longer. The majority voting method performed the best because it combined the knowledge

of multiple rules.

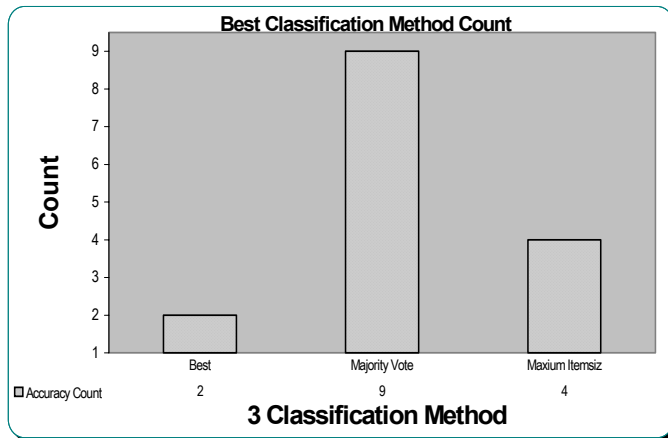


Figure 4 – Performance of Different Associative Classification Methods in 15 Stocks’ Intra-stock Mining

B. Performance of Associative Classification Methods in Inter-stock Mining

As the itemset size for inter-stock mining is always the same for the selected combination, i.e., if *n* stocks are selected, the itemset size is always equal to *n*, analysis of the associative classification methods has been focused on Majority Voting and Best Confidence only. That is, the Maximum Window Size method is not applicable here. Ten inter-stock mining cases were selected as follows.

Table VIII. Ten Inter-stock Mining Cases

Antecedents	To Predict (Consequence)
1. 0001 - Cheung Kong (Properties) & 0002 – CLP Holdings (Utilities)	0003 – HK & China Gas (Utilities)
2. 0002 – CLP Holdings (Utilities) & 0003 – HK & China Gas (Utilities)	0004 - Wharf Holdings (Conglomerates)
3. 0001 - Cheung Kong (Properties) & 0005 (Banks) – HSBC Holdings	0011 – Hang Seng Bank (Banks)
4. 0004 - Wharf Holdings (Utilities) & 0006 – HK Electric (Utilities)	0010 – Hang Lung Group (Properties)
5. 0008 – PCCW (Telecom) & 0012 - Henderson Land	0014 – Hysan Development (Properties)
6. 0016 – Sun Hung Kei Property (Properties) & 0017 - New World Development (Conglomerates) to predict 0019 - Swire Pacific (Conglomerates)	0008 – PCCW (Telecom) and 0012 - Henderson Land
7. 0330 - Esprit Holdings LTD (Textiles & Clothing) & 709 - Giordano International LTD (Textiles & Clothing)	0440 - DAH SING Financial Holdings LTD (Bank)
8. 0013 – Hutchison (Conglomerates) & 0023 – Bank of East Asia (Banks)	0005 – HSBC Holdings (Banks)
9. 0054 - Hopewell Holdings (Conglomerates) & 0101 - Hang Lung Properties (Properties)	0001 - Cheung Kong (Properties)
10. 0005 – HSBC Holdings (Banks) & 0008 – PCCW (Telecommunication)	0116 - Chow Sang Sang (Retailers)

Again, the Majority Voting method yields better prediction result. It is convincing because the result from the Best

Confidence method may be deteriorated by low support count.

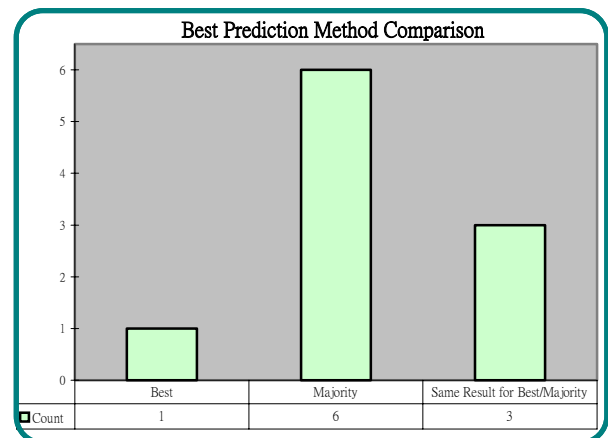


Figure 5 – Performance of Different Associative Classification Methods in 10 Cases’ Inter-stock Mining

C. Performance Analysis of Different Stock Sectors in Inter-stock Mining

Here, the inter-relationship among different sectors of stocks is investigated. They are detailed as follows.

Table IX. The Inter-relationship Among Different Sectors of Stocks

1. Mother/Sister Companies	: 0001 – Cheung Kong predicts 0013 – Hutchison
2. Same Sector - Properties	: 0012 – Henderson Land predicts 0016 – Sun Hung Kei Property
3. Same Sector - Utilities	: 0002 – CLP Holdings predicts 0003 – HK & China Gas
4. Same Sector - Banking	: 0005 – HSBC Holdings predicts 0011 – Hang Seng Bank
5. Same Sector - Manufactory	: 0330 – Esprit Holdings predicts 0709 – Giordano Int'l
6. No Special Relationship, but both are Blue Chips	: 0001 – Cheung Kong predicts 0002 – CLP Holdings 0001 - Cheung Kong predicts 0005 – HSBC Holdings 0001 - Cheung Kong predicts 0011 – Hang Seng Bank 0001 - Cheung Kong predicts 0008 – PCCW 0003 – HK& China Gas predicts 0004 - Wharf Holdings

As expected, Fig.6 shows that the mother/sister companies Stock 0001 and Stock 0013 yield the best inter-stock associative classification result. Here, the majority voting method was adopted. It can also be seen that the inter-relationship of stocks in the same industrial sectors is strong and has produced encouraging classification results, e.g., Properties Sector: Stock 0012 – Henderson Land and Stock 0016 – Sun Hung Kei Property, Public Affairs Sector: Stock 0002 – CLP Holdings and Stock 0003 – HK & China Gas, and Banking Sector: Stock 0005 – HSBC Holdings and Stock 0011 – Hang Seng. For the stocks that are blue chips, they have produced quite good classification result, but not as good as the mother/sister company pair nor the same Industrial Sector pairs.

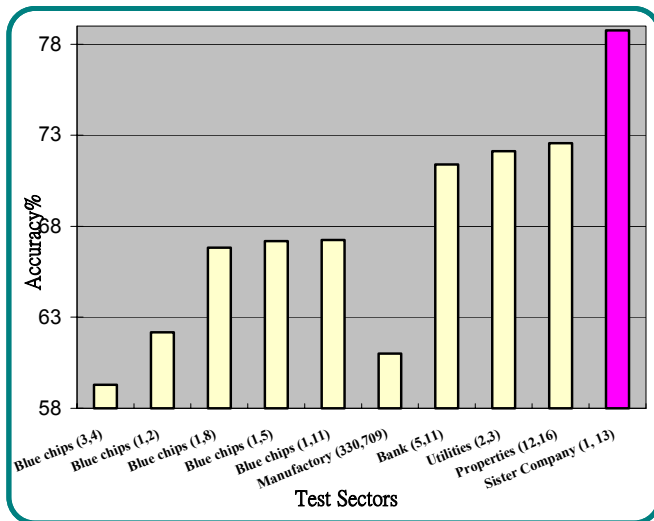


Figure 6 – Performance of Inter-stock mining for Different Sectors

VI. CONCLUSIONS

In this paper, the problem of pattern-based stock data mining is addressed and two mining approaches based on association analysis are proposed. They are: (i) intra-stock mining which focuses on finding frequently appearing patterns for a selected stock itself and (ii) inter-stock mining which finds frequently appearing inter-relationships among several stocks. We propose to make use of the sequential and non-sequential association rules mined to predict future stock price movements. Three associative classification methods, namely, best confidence, maximum window size and majority voting, are adopted to select appropriate association rule(s) and make prediction of the future stock price movements. In addition, we propose to transform the numeric stock prices into symbolic strings based on candlestick charting or technical analysis so that association rule mining can be applied. Various experiments have been conducted to evaluate the performance of our associative classification methods and the effectiveness of intra-stock mining and inter-stock mining. The proposed methods are efficient and quite effective in predicting the price movements as well as understanding the inter-relationship among stocks, say from the same sector and mother/sister companies.

In this paper, we only concentrate on the price information from stock data. More influential features can be considered in the numeric-to-symbolic conversion process and be used to generate the association rules. To name a few, the transaction volume can be considered for trap discovery and prevention because even though the rules mined are very interesting and accurate, one cannot buy or sell easily if the transaction volume is low. Moreover, the candle relationship like Candle Upper/Lower Tail, Candle Total Weight and Candle position compared with previous transaction time can be used to generate other useful features. Besides, an optimal or auto-adjust support count can be considered for further investigation such as to support different stock markets and improve the classification/prediction accuracy.

REFERENCES

1. J. M. Zytkow, "The KDD land of plenty," *AAAI Workshop Notes – Knowledge Discovery Databases*, Anaheim, CA, pp. iii-vi, July 14, 1991.
2. G. Qu, S. Hariri, & M. Yousif, "A new dependency and correlation analysis for features," *IEEE Trans. on Knowledge and Data Engineering*, vol.17, no.9 pp.1199–1207, Sept. 2005.
3. Kyoung-Jae Kim, "Artificial neural networks with feature transformation based on domain knowledge for the prediction of stock index features," *Intell. Sys. Acc. Fin. Mgm.*, vol.12, pp.167-176, 2004.
4. R. Agrawal, T. Imielinski & A. Swami, "Mining association rules between sets of items in large databases," *Proceedings of the ACM SIGMOD Conference on Management of Data*, Washington, D.C., pp. 207-216, May 1993.
5. J. Han and M. Kamber, *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2001.
6. Reuters, *An Introduction to Technical Analysis*. John Wiley & Sons (Asia) Pte Ltd., 1999.
7. G. L. Morris, *Candlestick Charting Explained*, CandlePower, 1992.
8. R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," *VLDB'94*, Santiago, Chile, pp.487-499, 1994.
9. R. Agrawal and R. Srikant, "Mining Sequence Pattern," IBM Research Report RJ9910, October 1994.
10. B. Liu, W. Hsu, and Y. Ma, "Integrating classification and association rule mining," *Proceedings of the KDD*, New York, pp.80-86, 1998.