COMP 4433

Take-home Examination: Part I

Inter- and Intra-Stock Data Mining and Clustering

Report

DING Yukuan
18082849D

1. Introduction

Data mining could be effective in extracting the trends of stocks as well as the correlations among stocks. In this project, we are provided with 6 stocks (code 1, 11, 293, 857, 13, 23). Using only the close price of these stocks in 1856 days, we formulate a general clustering problem, and solve the problem with inter- and intra-stock data mining under specific scenarios.

2. Problem Formulation

2.1 Scenario 1: *Reduction of Risk through Investment Combination*

In stock investments, the portfolio optimization theory suggests that the low correlations among properties would result in lower risk of deficit. Here, the term 'correlation' specifically refers to simultaneous movement of prices in same direction (e.g., when stock A's price grows/reduces, stock B's price also grows/reduces). When an investor following such optimization theory encounters a deficit in one of his/her property, the other properties would have lower chances to trigger loss at the same time. Considering such correlation, clustering stocks with high correlations could notify the investor to avoid choosing stocks in the same cluster, which indeed reduces risk of loss.

In this scenario, the problem can be formulated as:

- Extracting features from stocks' daily closing prices, which can represent the similarities on the trends of prices among stocks.
- Applying clustering model to cluster the stocks based on extracted features.

2.2 Scenario 2: *Range Reduction for Searching Related Industries*

For policy makers, stock market is a platform for observing the relation network of different companies and industries, in order to make macro regulation and control. When an industry A is related to industry B, they can be majorly in either competitive relation or cooperative relation, which can be represented by whether their stock price trends simultaneously move in same or opposite direction. In real world situation, the existing relation can be more complicated, where the stock price trend of a company may be related to the combined conditions of several other companies. Thus, the methodology described in *Scenario 1* is vulnerable in extracting complex relationships, and we derive the formulation of the problem as follows:

- Compute the correlation (can be non-linear) between combination of stocks (e.g., between stock 1 and stock 11, and between stock 1 and stock pair (293, 857)).

- Use the correlation to represent dissimilarity among stocks and among stock pairs, and apply cluster model to utilize such dissimilarity to do clustering.

3. Data Preprocessing

In this project, we mainly focus on the moving trend of closing prices. Based on the closing prices of a day P_i and of the day before P_{i-1}, we could derive the moving trend T_i of day i by:

$$T_i = f(x) = \begin{cases} UP, & v > 0 \; and \; v > \lambda \\ LEVEL, & |v| \leq \lambda \\ DOWN, & v < 0 \; and \; |v| > \lambda \end{cases}$$

, where $\lambda \epsilon R^*$ is a constant threshold value, and

$$v = \frac{P_i - P_{i-1}}{P_{i-1}}$$

In the stock prices, there can be missing values, for that the open date of the stocks may have difference. For intra-stock mining, since we are extracting timeline-related features inside each stock, the missing values can be ignored. However, when it comes to inter-stock mining, it has to be promised that all stocks have trading record for each timestamp, as the correlations among stocks are to be extracted at each timestamp. We adopt a simple but effective way, which drops the time period where at least one stock considered in this analysis has no trading record.

4. Problem Solution

4.1 Scenario 1

4.1.1 Inter-Stock Mining

After preprocessing, we have the moving trends of the six stocks from 2000/04/10 to 2007/02/21, denoted by $UP = 0, LEVEL = 1, DOWN = 2$. When considering inter-stock feature extraction, the raw sequence of trends of each stock is itself a feature that can represent the movement of the stock price, which is also frequently utilized by existing projects.

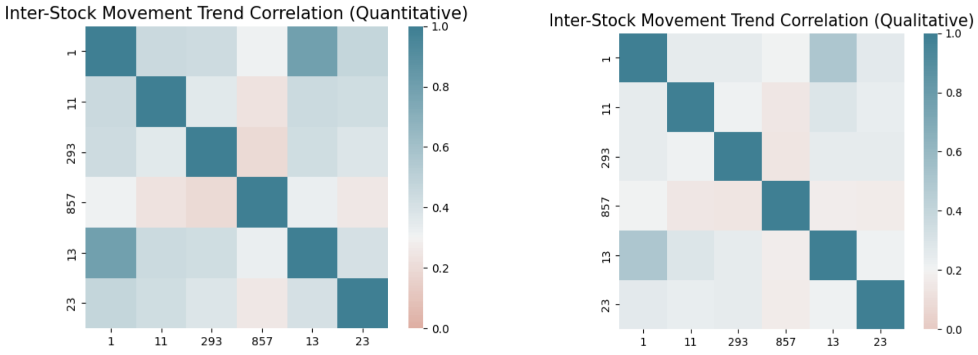

Figure 1: Quantitative (Based on daily change rate v) and Qualitative (Based on Up, Level, Down symbols) Correlation between Stocks' Trend Sequences on Whole Timeline

The correlations between pairs of stocks upon their whole trend sequence have shown unneglectable variances, which provides support for this methodology that it could differentiate the stocks' movements.

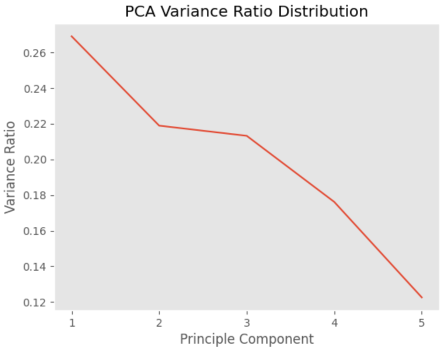

Figure 2: PCA Variance Ratio Distribution (sum = 0.99)

A pipeline including standardization and principal component analysis (PCA) has been applied to the sequences. For the simplicity of the project, we adopt K-means cluster model. The results under k=3 and k=4 is respectively:

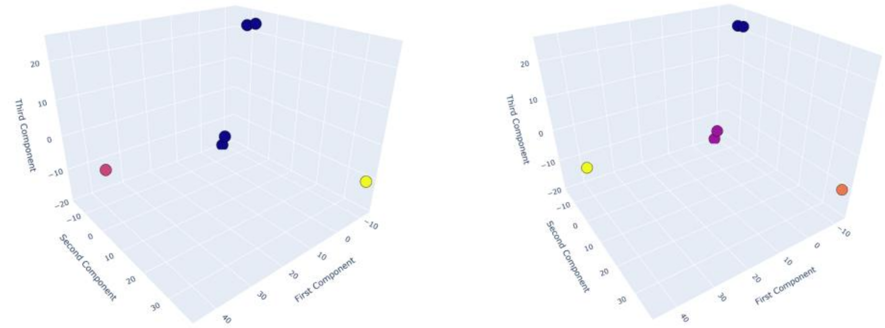

Figure 3: Scenario 1 Cluster Result 3D Plot with Inter-Stock Mining

K=3		K=4	
🟦	001, 011, 13, 23	🟦	011, 23
🟧	293	🟧	857
🟨	857	🟨	293
		🟪	001, 13

Table 1: Scenario 1 Cluster Result with Inter-Stock Mining

4.1.2 Intra-Stock Mining

In this section, we implement a modified Apriori Algorithm to conform to sequential data extraction [1].

We first introduce the term 'window'. In the movement sequence of a stock, we define n-window as the movement symbols in n consecutive transaction days. For example, the sequence $\{Up, Up, Down, Level, Up\}$ from $2001/01/01$ to $2001/01/05$ is a 5-window. What we want

In traditional Apriori Algorithm, the orders as well as repeats of items in each transaction is not considered. This modified algorithm follows the steps below:

1) Generate Frequent 1-itemset. This is just picking the symbols that satisfies minimum support, say $Up, Down\ Level$ (U, D, L in abbreviation)
2) By traditional algorithm, Candidate 2-itemset is generated by single-order join, which

results in $(U, D), (U, L), (D, L)$. However, inverted join and self-join are also required in this scenario where order has actual meaning and repeats of items are allowed. Thus, the candidate 2-itemset will also include $(D, U), (L, U), (D, L), (D, D), (U, L), (L, D)$. Items without enough support will be dropped to form frequent 2-itemset.

3) For n-itemset ($n \geq 3$), inverted join is not necessary, because the orders are generated in 2-items and can be brought into the following steps. While traditional Apriori Algorithm adopt left-most join, which combines the items with same leftmost values, we here adopt right-most join, since rightmost value is the most recent symbol that should be indication result of the generated rules. Thus, (U, D) and (L, D) would become (U, L, D). Self-join would also take place, which follows this right-most join strategy (e.g., (U, L) becomes (U, U, L))
4) For each item I in frequent m-itemset ($m \geq 2$), we formulate a rule

$$I[0:-1] => I[-1] \quad (support, confidence)$$

, where $I[0:-1]$ is the sequence representation of I except for its last (most recent) member.

Stock ID	D-D=>L (Support)	D-D=>L (Confidence)	D-D=>D (Support)	D-D=>D (Confidence)	U-D-D=>D (Support)	U-D-D=>D (Confidence)
1	0.064151	0.231068	0.154178	0.55534	0.042049	0.619048
11	0.0469	0.076855	0.513747	0.841873	0.03504	0.722222
293	0.058221	0.213439	0.158491	0.581028	0.033423	0.558559
857	0.044208	0.170626	0.148853	0.574514	0.031897	0.5
13	0.067385	0.204918	0.208625	0.634426	0.036119	0.57265
23	0.050674	0.129655	0.278167	0.711724	0.036119	0.587719

Table 2: A Part of the Mined Rules with Support Ratio and Confidence for Fitting K-means Model

Finally, for each of the stocks, several rules will be derived, with support and confidence calculated. These rules represent the price movement patterns in a period of time, and stocks with similar patterns are certainly low in dissimilarity. We then apply K-means upon the support ratio and confidence of the common rules shared by the six stocks. The results are as following.

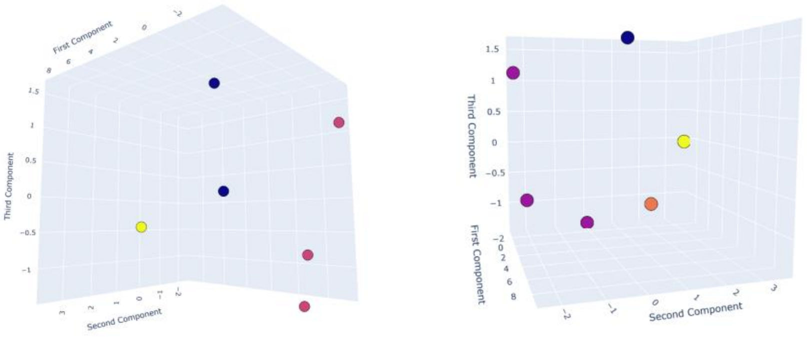

Figure 4: Scenario 1 Cluster Result 3D Plot with Intra-Stock Mining

K=3		K=4	
🟦	857, 23	🟦	23
🟥	001, 293, 13	🟧	857
🟨	011	🟨	011
		🟪	001, 293, 13

Table 3: Scenario 1 Cluster Result 3D Plot with Intra-Stock Mining

4.2 Scenario 2

In this scenario, we propose an inter-stock mining method that can derive further relations besides simultaneous movements towards same direction, among stocks and group of stocks.

The commence of this solution is based on Apriori Algorithm, where a transaction is the set of movement symbols of six stocks on a transaction day [1]. By Apriori Algorithm, we can finally get all frequent n-itemsets ($2 \leq n \leq 6$). However, there are some modifications:

- Each n-itemset will be attached to all size-n group of stocks (e.g., If we have only 3 stocks, an item in frequent 2-itemset {U D} is attached to (Stock 1, Stock 2), (Stock 1, Stock 3), and (Stock 2, Stock 3). Support and confidence are separately calculated in each of these attachments).
- Also, for example, pruning of a frequent 3-itemset attached to (Stock 1, Stock 2, Stock 3) will be based on checking only frequent 2-itemsets attached to (Stock 1, Stock 2), (Stock 1, Stock 3), and (Stock 2, Stock 3).

We then form a matrix like *Table 4*, where the column titles and row titles are the subsets of the 6 stocks (e.g., (1), (2), (3), (4), (5), (6), (1 2), (1 3), …, (2 3 4 5 6)). For a cell at row $(r_1, r_2, \dots r_p)$ and column $(c_1, c_2, \dots, c_q)$, it's value is calculated by:

- If any $r_i = c_j$, the value is NAN.
- If $p < q$, the value is NAN.
- Now we will have a set of stocks $(r_1, r_2, \dots r_p, c_1, c_2, \dots, c_q)$.We will then find a frequent $(p+q)$-itemset attached to this set of stocks. Assume an item is represented by $(a_1, a_2, \dots, a_p, b_1, b_2, \dots, b_q)$, where a_i is the symbol for r_i and b_j is the symbol for c_j. For each fixed $(a_1, a_2, \dots, a_p)$, when we propose a rule r: $(a_1, a_2, \dots, a_p) => (b_1, b_2, \dots, b_q)$, we can find the $(b_1, b_2, \dots, b_q)$ with highest confidence. Thus, we have a rule set $R = \{r_1, r_2, \dots\}$ that can predict symbols of $(c_1, c_2, \dots, c_q)$ when certain symbols of $(r_1, r_2, \dots r_p)$ are given. The value to fill in is the prediction accuracy, which can be conceptually considered to reflex the general (or average) confidence of the rule set R, as well as the *strength of the relation connecting these two stock groups.* In other words, this filled value is hypothetically a metric to measure dissimilarity.

	011	293	857	13	23	001	857,293	13,293
001	0.757	0.552	0.499	0.727	0.586			
011		0.556	0.494	0.619	0.621	0.587		
293	0.757		0.478	0.574	0.588	0.552		
857	0.770	0.517		0.566	0.594	0.524		
13	0.757	0.564	0.498		0.588	0.727		
23	0.757	0.547	0.491	0.578		0.587		
011, 001		0.568	0.509	0.731	0.635		0.327	0.458
293,001	0.757		0.507	0.727	0.625			
857,001	0.770	0.559		0.730	0.618			0.454

Table 4: Sample 'Relation Strength' Matrix

Hence, we use $(1 - matrix\, value)$ to form a dissimilarity matrix. With these dissimilarities, we conduct an approximate agglomerative hierarchical clustering, but using the pre-calculated dissimilarities between groups of stocks as cluster distances. The result dendrogram cluster is shown in *Figure 5*.

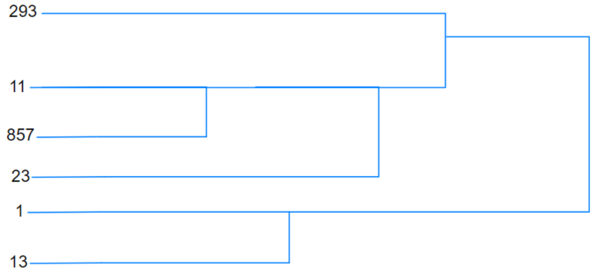

Figure 5: Scenario 2 Cluster Result dendrogram with Inter-Stock Mining

5. Conclusion and Future Works

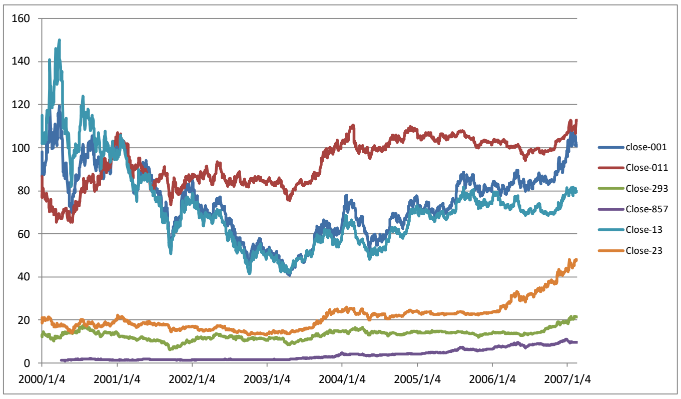

Figure 6: Trend of Close Prices of 6 Stocks

In this project, we implemented 2 inter-stock and 1 intra-stock mining approach to prepare data for clustering stocks under 2 different scenarios. Although we would not include systematic evaluation process in this project, we could have an approximate observation that all 3 solutions successfully found the visually similar stock pair (001, 13). In scenario 2, the clustering result strongly suggest the correlation between 011 and 857, which does not appear in both solutions in scenario 1, which may reflex the ability of the mining approach to further extract features that could represent complex relationship.

As all the three solutions are based on the preprocessing that only generates 3 symbols, in future we may add new symbols such as $Largely\_Up, Largely\_Down$ through introducing another threshold value. Also, in this project we are only dealing with closing price, while the other attributes may contain characteristic information rather than timeline sequential information, which could lead to faster and easier clustering. In addition, the clustering algorithm is another point for possible break through, where advanced clustering models can be adopted.

6. Acknowledgement

This work is derived from the classification/predication solution from [1].

7. References

[1] Ting, Jo, T. C. Fu, and Fu-lai Chung. "Mining of stock data: intra-and inter-stock pattern associative classification." *Threshold* 5.100 (2006): 5-99.