Pattern Discovery by Subsequence Time Series Clustering

ZHANG Caiqi 18085481d

The Hong Kong Polytechnic University

1 Introduction

In this report, we formulate a intra-stock clustering problem based on the given 6 Hong Kong stock data from 2001 to 2007. To be more specific, we apply the subsequence time series clustering to find the patterns in the stock data. The pipeline to solve the proposed problem is: we first generate the subsequences by sliding window, then prune the trivial match, and reduce the dimension of the subsequences for future clustering. We also conclude 6 the most frequent patterns from experiments.

2 Clustering Overview

Clustering is a very important task in data mining. The goal of clustering is usually to group similar objects together to discover some underlying rules. Methodologically, clustering algorithms can be classified as Centroid-based Clustering, Density-based Clustering, Distribution-based Clustering, Hierarchical Clustering, etc. In terms of the objects to be clustered, clustering analysis can be applied to images, static data, general sequences, and time series. The methods used are different for different clustering objects.

In this report, we will mainly focus on **time series analysis** and use six stocks of Hong Kong stocks from 2001 to 2007 as the data set.

3 Time Series Clustering Overview

With the advancement of data acquisition and data storage technology, we can keep data for a very long time for a certain variable. For example, stock market changes, exchange rate changes, blood pressure changes, heart rate changes, etc. Clustering for time series can help us to solve the following problems.

- Abnormality monitoring: for example, in blood pressure detection, if there is an abnormal change in blood pressure, an alert should be issued in time.

- Prediction and classification: the analysis of time series can help us to predict the subsequent changes. For example, we can predict the next trend of a variable based on its change in the recent period.
- Pattern recognition: time series usually have certain patterns of change, and these patterns can be summarized as several specific behavioral patterns. In practice, if we find that a stock is in a certain pattern, we can adjust accordingly to maximize the benefit.

This report will mainly focus on the application of **pattern discovery**.
According to [1], we can broadly classify time series clustering into three types.

- Whole time-series clustering: is the clustering of different complete time series. For example, in the Hong Kong stock dataset, the clustering of six stocks belongs to this type. Using whole time-series clustering can identify stocks with similar trends and apply to asset allocation, risk reduction, etc.
- Subsequence clustering: It is the task of dividing a long time series into multiple sub time series and clustering the sub series. Subsequence clustering can usually find the pattern within a certain series.
- Time point clustering: It is clustering oftime points based on a combination of their temporalproximity of time points and the similarity of the corre-sponding values

This report will focus on the analysis of **Subsequence clustering**.

4 Definition of Subsequence Time Series (STS) Clustering

According to [2], the definition of STS clustering can be described as follows: Let T be a real time series of length N:

$$T = [t_1, t_2, t_3, ..., t_N], t_i \in \mathbb{R}, i = 1...N$$

Sequences s_i are generated from t_i by applying a sliding window of width w:

$$s_n = [t_n, t_{n+1}, t_{n+2}, ..., t_{n+w-1}], n = 1...(N - w + 1)$$

Then the subsequences s_n can be clustered into several clusters in order to ind significant patterns in the series T.

5 The Importance and Application of STS Clustering

STS clustering can help us to find many interesting conclusions. Unlike inter clustering between different time series, all subsequences of STS clustering come from the same long time series, so it is easier to discover the structure and patterns within the sequences. [3]

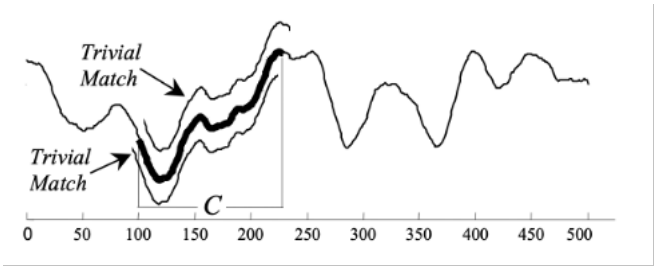

Figure 1: An example of the trivial match

also points out that pattern discovery is one of the most important applications of STS clustering. We can execute meaningful operations in the space of interest if we understand the data's structure. Predicting variable values, classifying previously unseen samples, and determining the likelihood of a given event are all examples of these procedures. As a result, discovering patterns is a necessary step in comprehending a given time series.

In the example of this report, we use STS clustering on pattern discovery of stock data. This has the following two most important benefits in reality: 1) it can help investors to identify the current stock patterns and adjust their investment strategies accordingly; 2) when the patterns are identified, the future trend of the stock can be predicted.

6 Statement of the Problem

According to the above definition and analysis, we can summarize our clustering problem as:

Conduct clustering for subsequences of six stocks' price in Hong Kong from 2001 to 2007, so as to discover the main patterns of stock price changes among them.

To solve this problem, we need to address the following three key questions. 1) How to generate the subsequences; 2) How to measure the similarity between two time series; 3) How to cluster the generated subsequences.

7 Problem Analysis

7.1 How to generate the subsequences?

In order to generate the subsequences from a long time series, a naive way is to use the sliding window, which is defined in Section 4. However, [4] points out that this approach has a very significant shortcoming because it will lead to many trivial match subsequences. According to [5], a trivial match, as shown in Figure 1, refers to two adjacent subsequences with a large degree of overlap. A large number of trivial matches will not be conducive to efficiently extracting useful information from the subsequences. Therefore, we must find a way to solve the problem of trivial match.

[6] proposed a threshold-free approach to improve the segmentation method for segmenting long stock time series into subsequences using sliding window. This approach leveraged the Perceptually Important Points (PIPs). The main purpose of the PIP is to reduce the

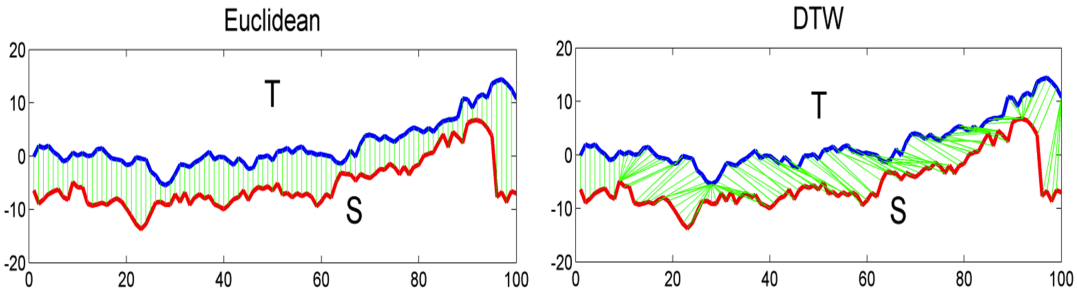

Figure 2: A comparison between Euclidean distance and DTW

dimensionality of the subseries and reflect the overall trend at the same time. The algorithm can be concluded in the following three steps: 1) Generate the subsequences using traditional sliding window method. 2) Filter the trivial match by PIP identification. 3) Compress the input patterns. [6] also experimentally demonstrated that this method not only finds the pattern in the time series well, but also can greatly reduce the running time by reducing the dimensionality of the data.

7.2 How to measure the similarity between two time series?

Since clustering is to combine similar time series together, how to measure the similarity between the series becomes the main problem we need to solve in the following. Metric distances is a simple and straightforward class of practices. The similarity of two series is quantified by measuring the distance between points in the time series. Common functions include: Euclidean Distance, Manhattan Distance, Maximum Distance, Minkowski Distance, etc.

However, the simple use of distance as a criterion has several problems: 1) it cannot identify the shape similarity, 2) it cannot reflect the similarity of the magnitude of the dynamic change of the trend, and 3) the calculation based on the point distance cannot reflect the difference of different frequencies (f).

7.2.1 Dynamic Time Warping (DTW)

The similarity computation is more robust using DTW. [7] argues that because it substitutes the one-to-one point comparison used in Euclidean distance with a many-to-one comparison, this approach may compare time series of diverse lengths. The key advantage of this distance measure is that it can identify comparable forms even if they have signal changes like shifting and/or scaling. A comparison between Euclidean distance and DTW is shown in Figure 2.

7.3 How to cluster the generated subsequences?

Many clustering methods can be applied to cluster the time series. According to the requirement of the project, we will simply adopt the k-means algorithm with some modification with respect to time series features. The use of other sophisticated clustering methods will be introduced in future work.

- Modification 1: Dynamic Time Warping (DTW) is used to measure the distance between time series.
- Modification 2: Cluster centroids are computed with respect to DTW. A centroid is the average sequence from a group of time series in DTW space.

8 Experiments

8.1 Configuration

In this section, a simple version of the above mentioned pattern discovery pipeline is implemented. For the sliding window size w, any positive relatively large integer is acceptable and we choose 49 here. There are around 1700 subsequences after applying the sliding window. For each subsequence of length 49, we apply PIP algorithm to reduce its dimension to only 9 points. We adopted the DTW and soft-DTW as the similarity metrics. The clustering is conducted by the **tslearn**, which is a python package that provides machine learning tools for the analysis of time series.

8.2 Effectiveness of PIP algorithm

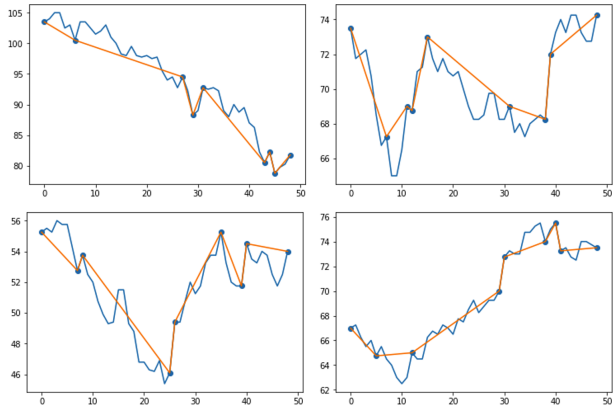

Figure 3: Blue lines show the time series before dimension reduction; orange lines indicate the trend after dimension reduction.

According to 3, before the dimensionality reduction, the original sequences are shown in blue, and the overall trend is not clear, which is not conducive to our analysis. After PIP processing (the orange line), the trend of the whole series is well reflected. At the same time, the time required for calculation was greatly reduced after dimensionality reduction. It was calculated experimentally that it took only 19 seconds to cluster the dimensionally reduced data while it took 118 seconds to cluster the original data.

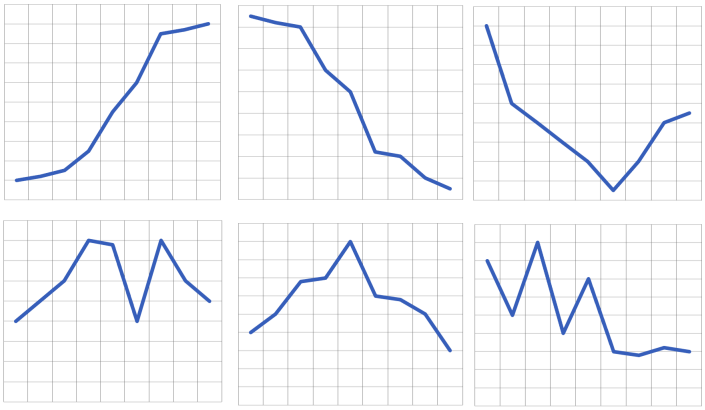

Figure 4: Main patterns found from the clustering result

8.3 Clustering Results

The final clustering results can be found in Figure 4. Figure 4 shows 6 the most frequent patterns found from the the stock data. We can find that some of them are very similar to the classical stock pattern (e.g. Head and Shoulder pattern), which proves the effectiveness of the overall clustering pipeline.

9 Future Work

- Variable length window: In this project, we only use a fixed window length. The setting of the window length is entirely based on experience and experimentation. Also, in real life, the window length involved in a pattern may not be fixed. In the future, we can attempt to optimize the algorithm by using a sliding window of variable length.
- Different Metrics: In this project, DTW and soft-DTW are used as similarity measures. But as introduced in S

  ection 7.2, this is not the only choice. In the future, we can try to use different metrics to compare their clustering performance.
- More clustering methods: In addition to the time series k-means method provided by tslearn, there are many other methods that can be used to cluster time series, such as Hierarchical clustering and Density-based clustering.

10 Conclusion

In this report, we formulate a intra-stock clustering problem based on the given 6 Hong Kong stock data from 2001 to 2007. We then apply the subsequence time series clustering to find the patterns in the stock data. A pipeline to solve the proposed problem is also introduced. Experiments have shown the effectiveness of the proposed solution.

Appendices

References

[1] S. Aghabozorgi, A. S. Shirkhorshidi, and T. Y. Wah, "Time-series clustering–a decade review," *Information Systems*, vol. 53, pp. 16–38, 2015.

[2] K. A. Peker, "Subsequence time series (sts) clustering techniques for meaningful pattern discovery," in *International Conference on Integration of Knowledge Intensive Multi-Agent Systems, 2005.* IEEE, 2005, pp. 360–365.

[3] S. Zolhavarieh, S. Aghabozorgi, and Y. W. Teh, "A review of subsequence time series clustering," *The Scientific World Journal*, vol. 2014, 2014.

[4] E. Keogh and J. Lin, "Clustering of time-series subsequences is meaningless: implications for previous and future research," *Knowledge and information systems*, vol. 8, no. 2, pp. 154–177, 2005.

[5] J. Lonardi and P. Patel, "Finding motifs in time series," in *Proc. of the 2nd Workshop on Temporal Data Mining*, 2002, pp. 53–68.

[6] T.-c. Fu, F.-l. Chung, V. Ng, and R. Luk, "Pattern discovery from stock time series using self-organizing maps," in *Workshop Notes of KDD2001 Workshop on Temporal Data Mining*, vol. 1. Citeseer, 2001.

[7] C. Cassisi, P. Montalto, M. Aliotta, A. Cannata, and A. Pulvirenti, "Similarity measures and dimensionality reduction techniques for time series data mining," *Advances in data mining knowledge discovery and applications'(InTech, Rijeka, Croatia, 2012,*, pp. 71–96, 2012.