Clustering Historical Stock Market Data to Diversify Portfolios and Improve Market Analysis

Chanakya Vishal KP
International Institute of Information Technology,
Hyderabad
20161116
chanakyavishal.k@students.iiit.ac.in

Anirudh Kovuru
International Institute of Information Technology,
Hyderabad
20161189
anirudh.kovuru@students.iiit.ac.in

ABSTRACT

The study aims to find an appropriate method to cluster the stock market so that an investor can diversify his stock portfolio. It can also help us understand the market situation better. The stock information obtained can act as a helping aid to new or even seasoned investors helping them maximize their returns. The stocks in each cluster would have a similar trend and hence investing in stocks in different clusters would enable minimal loss for the investor. We aim to use dynamic time warping as our similarity function and use an optimal clustering algorithm to create our clusters.

1 INTRODUCTION

Today, the stock market has become a hot investment for people in economic activities. But opportunities and risks both exist within the market, it is necessary to analyze the volatile characteristics of the stock market. There are many shortcomings in the traditional methods of stock analysis. Using our clustering algorithm, we can analyze various factors impacting the trend of stocks and their returns correctly and reveal the deep reasons that are hidden behind the data.

The application of clustering in diversification of a portfolio has many benefits which ensure that even if one cluster is failing the other clusters still keep the investor afloat. The main focus is to understand the time series data of the stock market and to find measures to gain information from the data available. For time series analysis we used a modification of the dynamic time warping algorithm as a similarity measure.

2 LITERATURE REVIEW AND RELATED WORK

Karsten Martiny proposed a method of Unsupervised Discovery using Candlestick Patterns[1]. The author proposed a method of forecasting security price movement using a candle stick model. The visualization technique along with the necessity for stock prices was mentioned in [8].

E. Keogh presented a lower-bounding measure for calculating a distance measure for univariate time series. He showed that his lower bounding measure is tighter than previously proposed measures[Yi et al. 1998, Kim et al. 2001]. The paper[2] deals with comparing the two models, one being the Hidden markov model and the other one Dynamic Time Warping to check for which one is better. A multivariate version of the LB Keogh Lower Bounding of DTW was proposed in [4].

The practical usage of various data mining techniques to cluster historical stock market data for Knowledge Discovery along with an approach was proposed in [6].

3 STOCKS DATA SET

We used the historical stock data of the S&P companies[1] collected over the past 5 years. The data set contained the following set of attributes:

(1) Date: The date under consideration
(2) Open: The opening price of a given stock on a given day
(3) High: The maximum value of the share during the day
(4) Low: The minimum value of the share during the day
(5) Close: The closing price of a given stock on a given day
(6) Volume: The volume of stock traded on that day
(7) Ticker-Name: The stock under consideration

We have data for over 470 stocks with each stock containing Open, High, Low, Close and Volume data for almost 1259 dates.

4 FEATURES SELECTION

We had to choose the feature set we needed for the clustering. Our goal while selecting the feature set was to make sure that the features could properly depict the trends of the stock.

4.1 Open and Close Values

: The difference between Open and Close values gives us an approximate trend of the stock on that particular day.

$$\frac{Open - Close}{Open}$$

We also had to factor in **normalization** of the values. This was needed as two stocks may have a similar trend but are in different scales. For example both the stocks of a top tier technology company and a start-up may have the same trend in stock price but the stock value of the top tier company may be four times higher than the start-up. If no normalization is done then these two similar stocks wouldn't end up in the same cluster.

4.2 Volume

: The volume of trade of a certain stock also plays a vital role in deciding which cluster it belongs to, as stocks with similar volume of trade are similar to each other. Including volume directly as a feature would prove to be non-optimal as the range of volume is very high, hence we divide the value of volume by the average

[1] https://www.kaggle.com/camnugent/sandp500

Date	Open	High	Low	Close	Volume	Ticker-Name
2013-02-08	15.07	15.12	14.63	14.75	8407500	AAL
2015-02-06	120.02	120.25	118.45	118.93	43706567	AAPL
2016-06-27	682.49	683.325	672.66	681.14	2919486	GOOGL
2016-06-28	691.37	692.7399	684.85	691.26	1912280	GOOGL
2016-06-29	694.26	699.5	692.6834	695.19	2156218	GOOGL
2016-06-30	697.65	703.77	694.9015	703.53	2112513	GOOGL
2016-07-01	705.1	712.53	703.73	710.25	1549160	GOOGL
2016-07-05	705.01	708.12	699.13	704.89	1422028	GOOGL

Table 1: The stock data set

volume of that particular stock throughout the date range.

$$\frac{volume(stock, date)}{\sum_{j=1}^{dates} volume_{stock,j}}$$

4.3 High value

$$\frac{High}{Open}$$

We normalized the High value with the opening price[1].

4.4 Low value

$$\frac{Low}{Open}$$

We normalized the Low value with the opening price[1].

5 DYNAMIC TIME WARPING

Dynamic time warping (DTW) is an algorithm for measuring similarity between two temporal sequences. It seeks to find the optimal match between two time series data sets.

The usage of this method instead of a more common distance measure like **Euclidean measure** is because of 2 reasons:

(1) A time series data can be of different lengths, which can't be processed by the Euclidean measure
(2) We also need to capture the long term stock similarity trend. For example two similar stocks may show similar trends but the changes may have a few days gap.

Distance measures like **cross-correlation** can't be used because there may be a difference in the rate of progress of the data, meaning that within a single time series, the rate at which progress is made can vary non-linearly.

5.1 Local distance Measure for the DTW

: In order to align time series, a distance measure, which allows the similarity assessment of positions in two time series, must be defined. We use a Euclidean measure for finding the distance.

$$Distance(q_x, c_y) = \sqrt{\sum_{i=1}^{p} (q_{x,i} - c_{y,i})^2}$$

where p is the number of dimensions which in this case is 4 and q_x and c_y are single data points of two time series q and c.

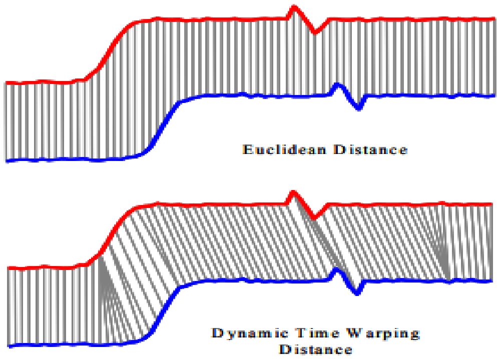

Figure 1: Note that while the two time series have an overall similar shape, they are not aligned in the time axis. Euclidean distance, which assumes the i^{th} point in one sequence is aligned with the i^{th} point in the other, will produce a pessimistic dissimilarity measure. The non-linear Dynamic Time Warped alignment allows a more intuitive distance measure to be calculated.[11]

5.2 LB Keogh Lower Bounding of DTW

: To make the Dynamic Time Warping algorithm more time efficient we can use **LB Keogh lower bounding**. The process of limiting how much the warping path may stray from the diagonal is known as **Lower bounding of a DTW**. Applying this optimization helps us prune out numerous expensive DTW computations. This reduces the time complexity from quadratic to linear.

Given a reach r and a time series q,

$$u_i = max(q_{i-r}, q_{i+r})$$

$$l_i = min(q_{i-r}, q_{i+r})$$

where for p dimensional data the bounds are,

$$u_i = (u_{i,1}, u_{i,2}, ...u_{i,p})$$

$$l_i = (l_{i,1}, l_{i,2}, ...l_{i,p})$$

Using the above we can write the LB Keogh measure between the time series c and the above mentioned q as,

$$LBKeogh(q,c) = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{p}\begin{cases}(c_{i,p}-u_{i,p})^2 & \text{if } \mathrm{c_{i,p}>u_{i,p}}\\(c_{i,p}-l_{i,p})^2 & \text{if } \mathrm{c_{i,p}<l_{i,p}}\\0 & \text{otherwise}\end{cases}}$$

6 VISUALIZATION TECHNIQUES

The stock variation over time can be visualized using a **Price candlestick chart**. The chart helps us capture the open, close, high and low prices or the **OHLC data** of the stock along with providing an inference of what the trends looked like on that day. They are also more visually appealing and easier to understand compared to regular bar charts.

In this chart, "candlesticks" are created using the open and closing prices. The opening and closing prices are plotted and a candlestick is extended towards the closing price from the opening price. If the candlestick extends downwards or the **closing price is lower than the opening price**, the candlestick is coloured red. On the other hand, if the candlestick extends upwards or the **closing price is greater than the opening price**, the candlestick is coloured green. This filled portion is referred to as the "body" of the candlestick. The long thin lines above and below the body represent the high/low range and are called "shadows" (also referred to as "wicks" and "tails").

Therefore, each candlestick provides an easy-to-decipher picture of price action. A trader can immediately compare the relationship between the open and close as well as the high and low. **The relationship between the open and close is considered vital information and forms the essence of candlesticks**.

Green candlesticks, where the close is greater than the open, indicate buying pressure. Red candlesticks, where the close is less than the open, indicate selling pressure. The chart also gives information about the magnitude of price variation on that day. Characteristics of the candle such as its length, colour and size of its shadows tell us a lot about the buying and selling sentiments for that on stock on that day.

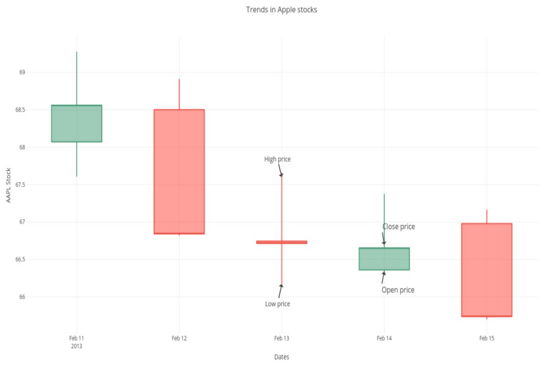

Figure 2: A candlestick chart showing price variation on a given day. Red indicates bearish while green indicates bullish markets.

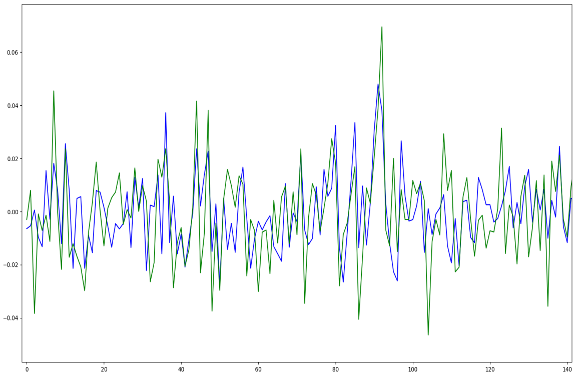

Figure 3: The above plot shows the variation of the $open - close/open$ **attribute within a time interval for two stocks TEL and PX which lie in the same cluster. Notice the similarity in the peaks of these two plots.**

7 CLUSTERING ALGORITHM

We decided to use the K-means clustering algorithm in order to generate the stock groups. We used K-means clustering because of the following reasons:

(1) Due to the large amount of data that we had to process. We had to process 470 stocks x 1259 dates coupled with the fact that we had to run Dynamic Time Warping which is also an O(n^2) algorithm. We lacked the resources to run the algorithm for all stocks until we could get meaningful results. As such running the algorithm took an entire day on our systems.

(2) Another important reason we had proposed the use of K-means was that the clusters formed would be tighter and this would lead to higher emphasis on the use of our distance measure.

That being said K-means does have its own fair share of problems which we had to take into account such as deciding the number of clusters, initialization, etc.

We used the LB Keogh Lower Bounding to prune out any unnecessary distance computations and used the Dynamic Time Warping algorithm to calculate the distance between two sequences. We ran the algorithm on our data set until it reached convergence, i.e., the mean values did not change significantly (or) the clustering assignments didn't undergo any change. For improving the speed of the algorithm we hyper threaded our clustering algorithm. More details on the implementation can be found from our publicly available source code[2].

8 EVALUATION CRITERIA

We evaluated our algorithm by using the silhouette measure[10][3] to intrinsically check the quality of our clustering. We found the silhouette measure to be the best for meaningful results at around 8 clusters, the average silhouette score being around 0.8932. The

[2] https://github.com/anirudhkovuru/Stock-cluster

[3] Implementation can be found in source code

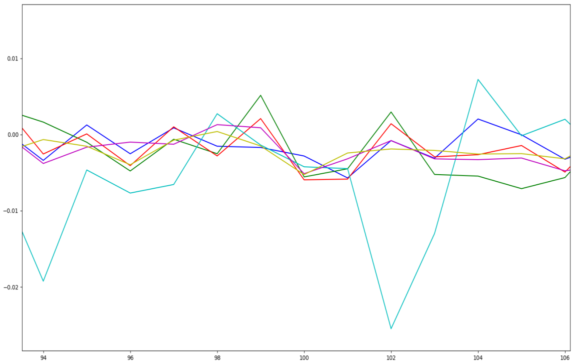

Figure 4: The above plot shows the variation of the $(open - close)/open$ attribute within a time interval for the different cluster centroids.

mean/centroid provide insight into the total nature of all the data of that cluster.

We visualized the means of each cluster by plotting them together. Due to the fact that the centroids would have normalized versions of the open, close, high and low values, we could not plot a candlestick graph. We, therefore, plotted any value against date for all the centroids to visualize the results.

We also found clusters with a maximum of 1 or 2 values within them. Due to this behaviour these stocks have been labeled as outliers but this also could be possible due to any form of bias in the chosen data set. Due to this the number of clusters we have computed has reduced to 6.

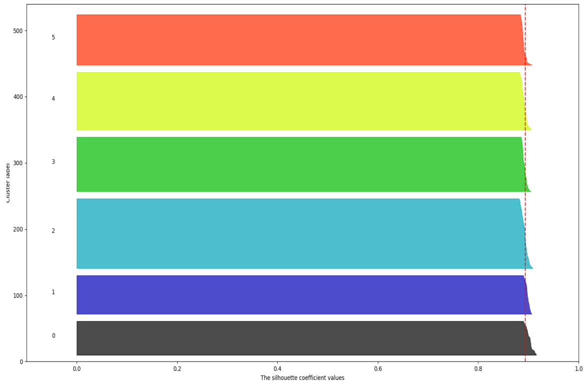

Figure 5: The above plot shows the silhouette score for each point grouped in clusters. The average silhouette score is represented by the red dotted line.

9 CONCLUSION

This paper presents an algorithm for time series clustering applied on the historical stock data. This algorithm improves market analyses by removing human bias, it also improves investor returns as more insight is available for making a well balanced portfolio. We extract and model the feature data based on the salient aspects of the stock market domain. We then cluster the time series data obtained using a variant of K-means where the distance measure is defined using Dynamic Time Warping. Due to the high dimensionality of time series data it is important to study the clustering of such kind of data deeply.

10 REFERENCES

(1) Karsten Martiny, "Unsupervised Discovery of Significant Candlestick Patterns for Forecasting Security Price Movements", Institute for Software Systems (STS), Hamburg University of Technology, Hamburg, Germany

(2) Tim Oates, Laura Firoiu and Paul R. Cohen: Clustering Time Series with Hidden Markov Models and Dynamic Time Warping Computer Science Department, LGRC University of Massachusetts

(3) Eamonn Keogh, Chotirat Ann Ratanamahatana, "Exact indexing of dynamic time warping", University of California Riverside

(4) Toni M. Rath and R. Manmatha, "Lower-Bounding of Dynamic Time Warping Distances for Multivariate Time Series", Multi-Media Indexing and Retrieval Group Center for Intelligent Information Retrieval University of Massachusetts

(5) Chotirat Ann Ratanamahatana, Eamonn Keogh, "Making Time-series Classification More Accurate Using Learned Constraints"

(6) Van-Dai Ta, Chuan-Ming Liu, "Stock Market Analysis Using Clustering Techniques: The Impact of Foreign Ownership on Stock Volatility in Vietnam"

(7) Nguyen Cong Long, Nawaporn Wisitpongphan, Phayung Meesad, "Clustering stock data for multi-objective portfolio optimization"

(8) Hans Salmonsson, "Finding Predictive Patterns in Historical Stock Data Using Self-Organizing Maps and Particle Swarm Optimization"

(9) Abhi Dattasharma, Praveen Kumar, Tripathi Sridhar, "Identifying Stock Similarity Based on Multi-event Episodes"

(10) Peter J. Rousseeuw, "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". Computational and Applied Mathematics.

(11) Young-Seon Jeong, Seong-Jun Kim, M.K. Jeong, "Automatic Identification of Defect Patterns in Semiconductor Wafer Maps Using Spatial Correlogram and Dynamic Time Warping"