Rattanakosin Journal of Science and Technology: RJST
2022, Volume 4 Issue 3: 23-39
*Research Article*
*Online ISSN: 2673-0553*

# Comparison of the classification efficiencies of K-means and Hierarchical clustering methods for candlestick component length on the world gold price candlestick chart

Sarinna Maplook[1*]

[1*]Statistics Program, Faculty of Science, Maejo University

[*]E-mail: sarinna@mju.ac.th

**Abstract**

In this study, the classification efficiencies of K-means and Hierarchical clustering methods for candlestick component length on the world gold price candlestick chart were compared with the objective to reduce the ambiguity in the identification of candlestick size. The data used in this study consisted of the opening price, closing price, highest trading price and lowest trading price from the daily gold price in the world market of the United States during January 2, 2012 to April 30, 2021 for a total of 2386 days. The data was divided into 2 groups: training data and testing data. The experiment consisted of the calculation of length, ratio, mean and standard deviation and standardization. The data was then classified using K-means and hierarchical clustering methods which it was found that K-means clustering method resulted in 5 clusters, while hierarchical clustering method resulted in 3 clusters. The classification efficiencies for each component of candlestick chart based on CCI values in both training data and testing data were all higher than 70 for both clustering methods. This indicated that the classification was quite effective. However, K-means clustering method was more effective compared to hierarchical clustering method for gold price data in the world market.

**Keywords:** candlestick, world gold price, K-means clustering, Hierarchical clustering, efficiency

## 1. Introduction

In the investment analysis in the gold market, in addition to considering economic fundamentals, trading timing is also important in investment. The role that gold plays in the economy and financial system is significant. Due to the way that gold prices fluctuate in the opposite direction from economic indices, it is a useful asset for successful diversification. It demonstrates that gold can maintain qualities that are unrelated to periods of market turbulence Anytime there is political unrest Gold is a secure investment to diversify your portfolio in times of financial uncertainty or high inflation because it is less volatile than other assets, especially those with low volatility. It can also boost investors' returns while lowering the total risk of the investment portfolio.

* Corresponding author, e-mail: sarinna@mju.ac.th
Received date: 3/ 08/ 2022 Revised date: 09/ 11/ 2022 Accepted date: 10/ 11/ 2022

Candlestick chart is another popular method for technical analysis as it can provide more details on the price data compared to the line charts. Candlestick chart reflects the trading behavior of the investors at a certain point of time for predicting the future price trends, whether short-term, medium-term or long-term trends. In considering the pattern of candlestick chart, the skilled investors will be able to immediately identify the pattern of the chart and then decide to buy or sell. But for unskilled investors, they may not be able to identify the patterns as good as it should be. (Farley, 2015)

Theory of Candlestick Analysis It was created by Honma Munehisa, also known as Sokyu Homma, a Japanese rice dealer who has been compiling daily rice price information for many years. People in the market study psychology Candlestick charts are currently frequently used to examine investor behavior when trading equities on the market. Technical analysis benefits It is merely a statistical data analysis. used to predict trade volume and prices. Popularity of candlestick charts is high. Because it can clearly and thoroughly convey information while also being simple to understand when viewed as a symbol, it is useful for analysis in determining market trends and directions. (Chen S, 2016)

Each candlestick consists of 4 components: opening price, closing price, highest trading price and lowest trading price in the specified period.

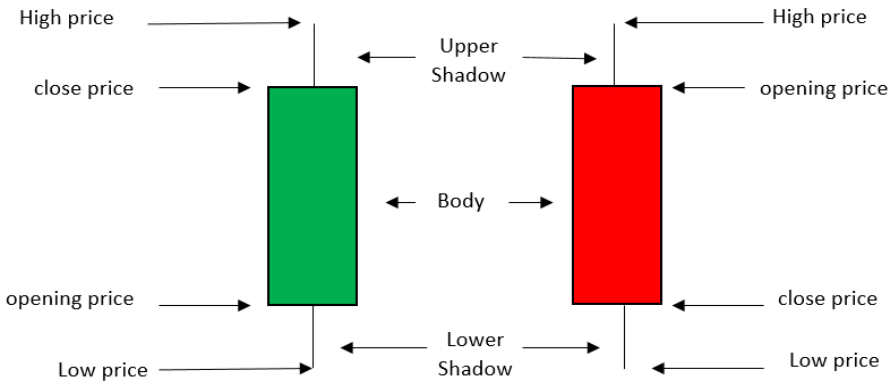


**Figure 1** Candlestick Pattern

It can be observed that the candlestick chart components which consist of sizes of upper shadow, body and lower shadow, are all important for the analysis of candlestick chart pattern.

In this study, the classification efficiencies of K-means and Hierarchical clustering methods for candlestick component length on the world gold price candlestick chart were compared with the objective to reduce the ambiguity in the identification of sizes on the candlestick chart. The results of the classification can be used in determination of candlestick chart pattern for the analysis of future gold price trends. To compare the classification efficiencies of K-means and hierarchical clustering methods for candlestick component length on the world gold price candlestick chart.

## 2. Methodology

In this study, the classification efficiencies of K-means and hierarchical clustering methods for candlestick component length on the world gold price candlestick chart were compared. The details of the classification methods used in this study are as follows:

### 2.1 Classification

2.1.1 K-Means clustering

K-means algorithm is one of the partitioning-based clustering algorithms. The general objective is to obtain the fixed number of partitions/clusters that minimize the sum of squared Euclidean distances between objects and cluster centroids. (Sudhir Singh, Dr. Nasib Singh., 2013) K-Means clustering is a popular technique for data classification. In this technique, the data is classified into K clusters as defined by user by determining the distance between each data point and the centroid of each cluster. The centroid of each cluster is an average of each attribute of the data in that cluster. The data classification procedure by K-Means clustering technique is as follows:

Step 1. Determine the centroid of each cluster by randomly assigning the cluster as the specified number.

Step 2. Calculate the distance between each data point and centroid of each cluster obtained from the previous step and reassign the data to be in the cluster near the centroid of each new cluster. In this study, Euclidean Distance formula was used and the data was classified to compatible with the nearest centroid as the following equation:

$$D_{ij} = \sqrt{\sum_{k=1}^{p} (X_{ik} - X_{jk})^2} \quad (1)$$

Step 3. Recalculate the centroid of each new cluster.

Step 4. Repeat Step 2-3 until all data points are in its original clusters or repeat for the specified cycles.

2.1.2 Hierarchical clustering

Hierarchical cluster analysis (HCA), also known as hierarchical clustering, is a popular method for cluster analysis in big data research and data mining aiming to establish a hierarchy of clusters. (Muntaner C, Chung H, Benach J, 2012) Hierarchical clustering is a method of cluster analysis with the following steps:

Step 1. Divide the data into n clusters of only 1 member.

Step 2. Merge 2 clusters to form a single cluster by considering the distance and using the specified criteria.

Step 3. Consider whether the 3rd cluster should be merged with the first 2 clusters or merged with the new cluster by considering the distance and using the specified criteria.

Step 4. Consider whether the 3rd cluster should be merged with the first 2 clusters or merged with another new cluster by considering the distance and using the specified criteria as Step 2.

Step 5. Repeat Step 3 until only a single cluster remains.

Step 6. Plot a Dendrogram and define the distance and then draw the line through the dendrogram to divide the data. (R.A. Johnson and D.W Wichern, 2018)

**2.2 Data**

The data used in this study was the Primary data is daily gold price data in the world market of the United States during January 2, 2012 to April 30, 2021 for a total of 2387 days[1]. The data was divided into 2 groups: training data and testing data with the following steps:

2.2.1 The classification of the length of each component on the candlestick chart is as follows:

Step 1. Calculate the distance of each component of the candlestick by dividing it into 3 parts: upper shadow (L_US), body (L_B) and lower shadow (L_LS) which calculated from

the opening price, closing price, highest trading price and lowest trading price with the following formula:

$$L\_US_i = HIGH_i - \text{Max}(OPEN_i, CLOSE_i), \quad (2)$$

$$L\_B_i = \|CLOSE_i - HIGH_i\|, \quad (3)$$

$$L\_LS_i = \text{Min}(OPEN_i, CLOSE_i) - LOW_i \quad (4)$$

where, HIGH is the highest daily gold price

OPEN is the daily opening gold price

CLOSE is the daily closing gold price

LOW is the lowest daily gold price

Step 2. Calculate the ratio of each component length of the candlestick using the following formulas:

$$r\_US_i = \frac{L\_US_i}{L\_C_i}, \quad (5)$$

$$r\_B_i = \frac{L\_B_i}{L\_C_i}, \quad (6)$$

$$r\_LS_i = \frac{L\_LS_i}{L\_C_i} \quad (7)$$

where, $L\_C_i = L\_US_i + L\_B_i + L\_LS_i$.

Step 3. Since the data must be standardized, it is therefore important to calculate mean and standard deviation of each component of the candlestick in the separation by using the following formulas:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \quad (8)$$

$$S.D = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}. \quad (9)$$

where, $x$ is the daily price of gold

$\bar{x}$ is the average price of gold

n is the days of gold price

Step 4. Calculate the adjusted value of each component using the standardization method so that the data is normally distributed by using the following formula:

$$Z = \frac{x - \bar{x}}{S.D} \quad (10)$$

2.2.2 Classify each component of the candlestick using K-means and hierarchical clustering methods.

2.2.3 Compare the data in percentage from the classification of each component of the candlestick chart.

### 2.3 Efficiency comparison

The classification efficiencies in this study were measured using the correctly clustered instances (CCI) between the clusters derived from the observed data and the clusters derived from K-means clustering and Hierarchical clustering methods. CCI can be calculated as follows:

$$CCI = \frac{\sum_{i=1}^{k} m_i}{n} x100 , \quad (11)$$

where, $m_i$ is the number of data points clustered by the clustering technique that match with the original data used in that cluster.

n is the total number of data points in the data set.

## 3. Results

The mean and standard deviation of candlestick components which was used in the standardization to adjust the ratio of each candlestick component size before the clustering is shown in Table 1

**Table 1** Mean and Standard deviation of candlestick components.

| LOWER | | BODY | | UPPER | |
|---|---|---|---|---|---|
| mean | S. D | mean | S. D | mean | S. D |
| 0.28342 | 0.18529 | 0.43571 | 0.25162 | 0.28012 | 0.19634 |

The classification of the candlestick components using K-means clustering method resulted in 5 clusters as follows:

**Table 2** Mean of the classification results of the candlestick components using K-means clustering method.

| Cluster | Open Price | High Price | Low Price | Close Price |
|---|---|---|---|---|
| 1 | 1542.996 | 1553.411 | 1532.229 | 1543.268 |
| 2 | 1304.894 | 1313.203 | 1296.77 | 1304.977 |
| 3 | 1896.497 | 1909.16 | 1881.297 | 1895.55 |
| 4 | 1709.836 | 1720.999 | 1698.11 | 1710.442 |
| 5 | 1186.152 | 1194.153 | 1178.236 | 1186.198 |

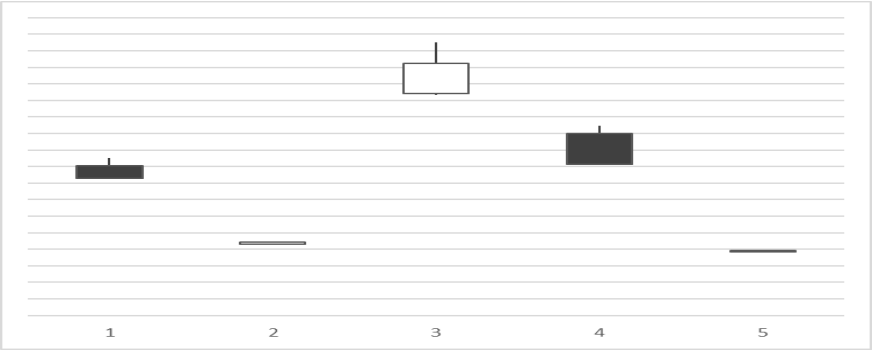


**Figure 2** Candlestick Pattern using K-means clustering method

The classification of the candlestick components using hierarchical clustering method resulted in 3 clusters as follows:

**Table 3** Mean of the classification results of the candlestick components using hierarchical clustering method.

| Cluster | Open Price | High Price | Low Price | Close Price |
|---|---|---|---|---|
| 1 | 1687.897 | 1699.045 | 1675.902 | 1688.148 |
| 2 | 1257.852 | 1266.071 | 1249.769 | 1257.899 |

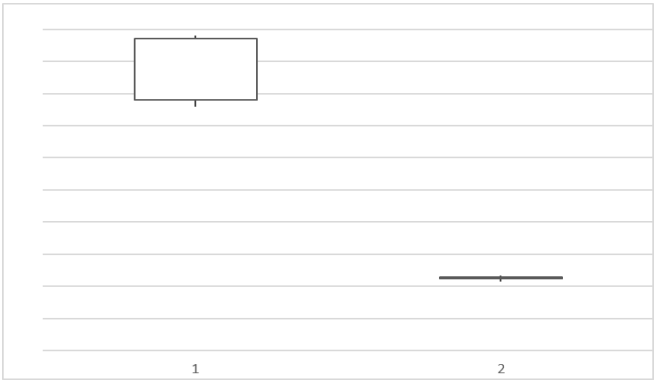


**Figure 3** Candlestick Pattern using Hierarchical clustering method

The results of the calculation of classification efficiencies (CCI values) in both clustering methods for each candlestick component are as follows:

**Table 4** CCI values in K-means and Hierarchical clustering methods.

| method | data | US | B | LS |
|---|---|---|---|---|
| K-means | training | 88.52 | 82.21 | 75.62 |
| | testing | 85.25 | 84.10 | 79.36 |
| Hierarchical | training | 71.35 | 73.34 | 72.37 |
| | testing | 73.25 | 72.65 | 78.29 |

This table shows the classification efficiencies for each component of the candlestick chart in form of CCI values for both training data and testing data. It can be observed that all values obtained from both methods were higher than 70, which indicated that the calculated centroids could be used to effectively classify the candlestick components for both training data and testing data. However, K-means clustering method was more effective than the hierarchical clustering method for gold price data in the world market.

## 4. Conclusion and Discussion

In this study, the classification efficiencies of K-means and Hierarchical clustering methods for candlestick component length on the world gold price candlestick chart were compared with the objective to reduce the ambiguity in the identification of candlestick size. The data used in this study consisted of the opening price, closing price, highest trading price and lowest trading price from the daily gold price in the world market of the United States during January 2, 2012 to April 30, 2021 for a total of 2386 days. The data was divided into 2 groups: training data and testing data. The experiment consisted of the calculation of length, ratio, mean and standard deviation and standardization. The data was then classified using K-means and hierarchical clustering methods which it was found that K-means clustering

method resulted in 5 clusters, while hierarchical clustering method resulted in 2 clusters. The classification efficiencies for each component of the candlestick using CCI in both training data and testing data were all higher than 70 for both clustering methods. This indicated that the classification was quite effective. However, K-means clustering was more effective compared to hierarchical clustering for gold price data in the world market. Therefore, the results of the classification may be used in the determination of candlestick chart pattern for the analysis of future gold price trends or other stock prices based on the analysis of candlestick chart patterns. This research has suggested that other clustering methods such as dynamic clustering or 2-step clustering should be also used for the better comparison of the classification efficiencies. And more factors such as economic and social factors should also be considered while using the gold price in the world market of the United State for more effective analysis.

## 5. References


[1] Chen S., Bao S., Zhou Y. (2016). The predictive power of Japanese candlestick charting in Chinese stock market. Physica A: Statistical Mechanics and Its Applications, 457, 148-165.

[2] Dubey A. K., Gupta U. and Jain S. (2018). Comparative Study of K-means and Fuzzy C-means Algorithms on The Breast Cancer Data Int. J. Adv. Sci. Eng. Inf. Technol. 8 18-29

[3] Fahim A. M., Salem A. M., Torkey F. A. and Ramadan M. A. (2006). An efficient enhanced k-means clustering algorithm, Journal of Zhejiang University Science A., pp. 1626–1633, 2006

[4] Farley, A. (2015, September 22). The 5 Most Powerful Candlestick Patterns. Retrieved April 3, 2018, from https://www.investopedia.com/articles/active-trading/092315/ 5-most-powerful-candlestick-patterns.asp

[5] Gradojevic, N., & Gençay, R. (2011). Financial Applications of Nonextensive Entropy [Applications Corner]. IEEE Signal Processing Magazine, 28(5), 116–141. https://doi.org/10.1109/MSP.2011.941843

[6] Leon, C.-H., WenSung, L., & Liu, C. A. (2005). Candlestick Tutor: an intelligent tool for investment knowledge learning and sharing. Fifth IEEE International Conference on Advanced Learning Technologies (ICALT'05) (p. 238–240).

[7] Muntaner C, Chung H, Benach J, et al. (2012). Hierarchical cluster analysis of labour market regulations and population health: a taxonomy of low- and middle-income countries. BMC Public Health.

[8] Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65.

[9] R.A. Johnson and D.W Wichern. (2018). Applied Multivariate Statistical Analysis.

[10] Sangsawad S. (2018). Candlestick Length Feature Clustering using K-Means Approach for Price Directions Analysis Case Study: SETHD index

[11] Sudhir Singh, Dr. Nasib Singh. (2013). Gill, Comparative Study of Different Data Mining Techniques: A Review, www.ijltemas.in, Volume II, Issue IV, APRIL 2013 IJLTEMAS ISSN 2278 – 2540.