THE APPLICATION OF K-MEANS ALGORITHM FOR LQ45 INDEX ON INDONESIA STOCK EXCHANGE

A. Raharto Condrobimo[1]; Albert V. Dian Sano[2]; Hendro Nindito[3]

[1,2,3] Information Systems Department, School of Information Systems, Bina Nusantara University
Jln. K.H. Syahdan No 9, Palmerah, Jakarta Barat, 11480
[1]condrobimo@binus.ac.id; [2]albert_vds@yahoo.com; [3]Hendro.nindito@binus.ac.id

ABSTRACT

The objective of this study is to apply cluster analysis or also known as clustering on stocks data listed in LQ45 index at Indonesia Stock Exchange. The problem is that traders need a tool to speed up decision-making process in buying, selling and holding their stocks.The method used in this cluster analysis is k-means algorithm. The data used in this study were taken from Indonesia Stock Exchange. Cluster analysis in this study took data's characteristics such as stocks volume and value. Results of cluster analysis were presented in the form of grouping of clusters' members visually. Therefore, this cluster analysis in this study could be used to identify more quickly and efficiently about the members of each cluster of LQ45 index. The results of such identification can be used by beginner-level investors who have started interest in stock investment to help make decision on stocks trading.

Keywords: *blue chip stock, data mining, k-means, clustering*

INTRODUCTION

Stock market development has been the subject of intensive theoretical and empirical studies. More recently, the emphasis has increasingly shifted to stock market indexes and the effect of stock markets on economic development (Athanasios & Antonius, 2012).

A share or a stock is the ownership relationship between the company and the shareholder or stockholder. Based on the classification, there are two types of stocks, that is, preferred stocks and common stocks. Preferred stocks are stocks that have special rights in a company (for example: the distribution of company profits received beforehand than the other stock owners) while the common stocks are stocks that do not have more rights in addition to the general right of obtaining profit sharing in accordance with the schedule of distribution of profits which will be convened in the Annual General Meeting of Stockholders (AGM). Common stocks (hereinafter referred to stocks) have advantages over special interests that can be transferred freely to other parties so that they can be traded in a market called stock.

Today Indonesia has only one stock market or stock exchange, that is, the Indonesian Stock Exchange (IDX). IDX provides mechanisms in selling and buying stocks for public owned companies listed in the IDX. Perseroan Terbatas (PT) is a legal entity to run a business which consists of capital stocks, which is a part owner of the shares they own. Public PT is a company with a limited liability company as well as the status of a public company (Go Public).

Share is a major product in the capital market instruments transacted. There are several derivatives arising from transactions that occur due to the stock exchanges. There are two ways to invest in stocks, first is buying and storing these shares so that the gain distribution of profits

(dividends) and second is buying and selling back shares so as to benefit from the difference between the buying and selling value (capital gain). Buying stocks in general can be done through two ways, bought during a stocks will rise and begin at its Initial Public Offering (IPO) and purchased through the secondary market that we are familiar with the stock market.

A few years ago the notion of shares was an investment only for the upper class. However, since the era of online trading increased where transactions could use online networking internet, stock transaction has increasingly shifted into an investment option for many people. It's because the minimum initial deposit today is more affordable.

With the more easily to get started investing in shares in the capital market, it is not only necessary to prepare the funds, but also requires a knowledge so that we can analyze the market situation at the time. Transaction in the stock market is actually the same as if we want to trade as usual. It required a skill in analyzing the current trends in order to trade in goods that we still exist and must be purchased by buyer's profit situation. To be able to analyze the market need, a sufficient education is needed so that ultimately have an analysis of its own. Currently it is not a bit of market participants who do not have sufficient knowledge, not even know yet, have already participated in the transaction market.

In a normal market situation and market environment which tends to have strengthened due to the state of the economy and strong corporate fundamentals, all market participants are capable being in a safe zone. However, in an downward moving markets just as what happened to 2008 as we faced together, market moved in any unpredictable direction, and it could drive investors just to follow the crowd, or follow gossip, and could get caught in a loss because the market moved to unwanted direction. Often a recommendation given by someone would work with the other way around, because it is related to the interests and the responses of the people. By being able to analyze independently, we are expected to be investor who are not easily affected by misleading information at that time.

To narrow the withdrawal of shares for approximately 500 stocks listed on our exchanges, we concentrate on stocks that are listed in LQ45 Index. LQ45 is a row of 45 stocks which are stocks with the most transaction in Indonesia Stock Exchange. That is why it is called LQ45 (Liquid 45).

What about the blue-chip stocks? There is no formal form for Blue Chips definitions on this day, even today this term become more common, therefore we do not provide a list based on LQ45, IDX. (n.d.).Why only those shares? We position ourselves all this time in a state of learning on the state of the market, and the stocks included in the index LQ45 are chosen as liquid stocks within the meaning of actively traded to keep us stuck in the second tier stocks that are sometimes played are very profitable and then hibernate in a long period of time making them hard to sell. In order to avoid a lot of things like that, we try to adapt to the index which is relatively safer for the transaction.

LQ45 Index is a market capitalization of the most 45 liquid stocks and has a large of capitalization. It is an indicator of liquidation. LQ45, using the 45 stocks are selected based on liquidity of stock trading and adjusted every six months (every early February and August). Thus the stocks contained in the index will always change.

Some of criteria in determining if an issuer can be included in LQ45 index are consisted of two criteria. The first criteria are: (1) being in the TOP 95% of the total average - the annual average value of share transactions in the regular market. (2) Being in the TOP 90% of the average - the annual average market capitalization. The second criteria: (1) it is the highest order which represents the sector in the Indonesia Stock Exchange (IDX) industrial classifications according to its market capitalization. (2) It is the highest order based on the frequency of transactions.

LQ45 index consists of 45 stocks that have been chosen through a variety of selection criteria, which will consist of stocks with liquidity and high market capitalization. Shares in LQ45 index must meet the selection criteria and pass the following key: (1) being in the top 60 of the total share transactions in the regular market (the average value of transactions during the last 12 months). (2) Ranking based on market capitalization (average market capitalization during the last 12 months). (3) It has been listed on the JSE at least 3 months. (4) The financial position of the company and its growth prospects, the frequency and number of trading days of regular market transactions.

Shares included in LQ45 continue to be monitored and will be held every six months review (early February and August). If there are shares that have not entered the criteria, it will be replaced with other shares that qualify. Selection process of shares LQ45 have to be reasonable, therefore Indonesia Stock Exchange has advisory committee consisting of experts in BAPEPAM, Universities, and professionals in the capital market.

The factors that play a role in the movement of LQ45are: (1) Indonesia Interest Rate as the benchmark of portfolio investment in Indonesia's financial markets. (2) The level of investor tolerance for risk. (3) Index mover stocks which in fact are large market capitalization stocks on IDX.

Factors that influence the rise of LQ45: (1) the strengthening of global and regional markets following a drop in world crude oil prices, and (2) the strengthening of the Indonesia currency exchange rate that can lift LQ45 to the positive zone.

The purpose of LQ45 is complementary for Composite Stock Price Index and in particular provides an objective and reliable tool for financial analysis, fund managers, investors and other capital market observers to monitor the price movements of stocks that are actively traded.

"We are living in the information age" is a saying popular; however, we are living in an era of data. The data in terabytes or petabytes poured into our computer network, worldwide web (www), and various data storage devices each day ranging from world business, community, science and engineering, medicine, and almost every other aspect of daily life. The explosive growth of the volume of existing data is the result of the process of computerization of our society and the rapid development of various devices the collection and storage of data which is terrific (Han and Kamber, 2012).

The explosive growth of data and widely available really make us aware that we are in the era of data. Various reliable and versatile tools are needed to automatically reveal valuable information from the large-volume data and transform it into the organized knowledge. This need has led to the birth of data mining. The field is still young, dynamic and promising. Data mining has been and will continue to make great strides in our journey from the era of data into the information age to come (Han & Kamber, 2012).

Data mining is the process of finding previously unknown patterns and trends in databases and using that information to build predictive models. Data mining provides a set of tools and techniques that can be applied to this processed data to discover hidden patterns and also provides healthcare professionals an additional source of knowledge for making decisions (Hossain,*et al.*, 2013)

Data mining is a fun way to extract various kinds of patterns, which presents knowledge implicitly stored in large datasets and focuses on matters related to its feasibility, usefulness, effectiveness and scalability. Data mining can also be seen as a very important step in the process to find knowledge. Data is normally done through a pre-process data cleansing, data integration, selection and transformation of data and prepared for mining. Data mining can also be done on different types of databases and data storage, but the type of pattern is found determined by different

types of functionality mining data such as descriptions, association, correlation analysis, classification, prediction, analysis of clusters, and so on (Tajunisha, 2010).

The concept of data mining, involves three steps i.e., capturing and storing the data, converting the raw data into information and converting the information into knowledge. Data in this context comprise all the raw material that an institution collects via normal operation. Capturing and storing the data is the first phase that is the process of applying mathematical and statistical formulas to "mine" the data warehouse (Kumar & Ramaswami, 2011).

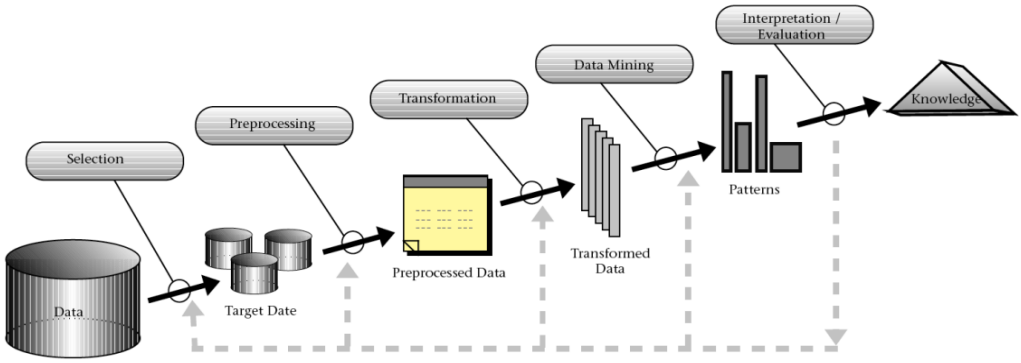

Figure 1 Data mining and knowledge discovery process of Database
(Sources: Fayyad, *et. al* in Silwattananusarn, 2012)

Based on Figure 1 above, the knowledge discovery process consists of several sequential and iterative methods such as the following (Fayyad, *et. al*, Han & Kamber, in Silwattananusarn, 2012): (1) selection: choosing relevant data to the task of a database analyst. (2) Preprocessing: deletingthe invalid data and inconsistent data; combining multiple sources of data. (3) Transformation: transforming the data into a suitable form to perform data mining. (4) Data Mining: choosing the data mining algorithm that matches the nature pattern of the data; extracting various data patterns. (5) Interpretation / evaluation: interpreting various patterns into knowledge by eliminating irrelevant various patterns and the same pattern and repetitive; translating a variety of patterns useful in terms that could be understood by ordinary people.

Clustering is an important method in data warehousing and data mining. It groups similar object together in a cluster (or clusters) and dissimilar object in other cluster (or clusters) or remove from the clustering process. However, there are some special requirements for search results clustering algorithms, two of which most important are, clustering performance and meaningful cluster description (Gothai & Balasubramanie, 2012).

Cluster analysis can also be called as clustering is the process of dividing a set of data objects (or object of observation) into several subsets. Each of these subsets is a cluster, such that the objects in a cluster are the objects that are similar to each other, but very different from the objects that are in another cluster. A set of clusters resulting from the cluster analysis such as clustering can be referred to clustering (Han & Kamber, 2012).

Cluster analysis offers a useful way to organize and present a complex dataset (Wang & Song, 2011). Analysis of the cluster can be regarded as the most popular techniques and foremost to solve problems that are unsupervised learning or undirected or unsupervised learning process. So each technique is used to solve problems. Certainly, a way of dealing with the structure of the data that has not been labeled will be found (Tayal & Raghuwanshi, 2011).

One important component of the clustering algorithm is a measure of the distance between data points. If a component of the vector sample data is in the same physical unit, then it is more likely that the simple Euclidean distance metric is sufficient to classify the data instants that are similar to each other. Tayal and Raghuwanshi (2011) stated that the distance between the two groups can be measured by (1) Euclidian and (2) City Block or Manhattan.

In addition to the similarity and dissimilarity of the two types of measurement above, some of the other measurements are shown in Table 2 below (Rui & Donald, 2005).

Table 2 Size of Similarity and Dissimilarity for Quantitative Variables (Rui & Donald, 2005)

Measures	Forms
Minkowski distance	$d(x,y)=\left[\sum_{i=1}^{p}\lvert x_i - y_i\rvert^m\right]^{1/m}$
Euclidean distance	$D_{ij}=\sqrt{\sum_{k=1}^{p}(X_{ik}-X_{jk})^2}.$
City-block distance	$D_{ij}=\sum_{k=1}^{p}\lvert X_{ik}-X_{jk}\rvert$
Mahalanobis distance	$D_{ij}=(X_i-X_j)^t S^{-1}(X_i-X_j)$, where S is the within group covarience matrix

K-means is one of the simplest undirected/no supervised (unsupervised) learning algorithms used to solve various problems of the grouping. The procedure is by applying a simple and easy way to classify data that has been given into a number of predefined clusters (such as clusters k) (Tayal & Raghuwanshi, 2011).

K-means algorithm will define the midpoint of the cluster from the average value of the points in the cluster. Steps in k -means algorithm can be explained as follows. First, the algorithm will select k (central cluster) randomly from various objects in D (dataset), which respectively represent the center of the cluster at the beginning or the first time. For any other object, each object is assigned or grouped into clusters that are the most similar or close based on the Euclidean distance between the object and the center of cluster.

K-means algorithm then iterates to improve or increase the separate distances or similarities in the cluster. For each cluster, this algorithm will calculate a new average using the objects that are grouped into a cluster in the previous iteration. All objects will then be regrouped by using the average of the newly updated as new cluster center. The iterations will continue until it reaches a stable grouping, which means that the clusters formed in the latest iteration is the same as the clusters formed in the previous iteration. K-means clustering procedure is generally summarized in Figure 2 below (Han & Kamber, 2012).

Algorithm: *k*-means. The k-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

Input:

- *k*: the number of clusters,
- D: a data set containing n objects.

Output: A set of k clusters.

Method:

1) Arbitrarily choose k objects from D as the initial cluster centers;
2) **Repeat**
3) (re)assign each object to the cluster to which the object is the most similar,
4) based on the mean value of the objects in the cluster;
5) Update the cluster means, that is, calculate the mean value of the objects for
6) Each clusters;
7) **Until** no changes;

Figure 2 Summary of Procedure Algorithm K-means (Han & Kamber, 2012)

As with any other algorithms, k-means also has some advantages and disadvantages. Here are the advantages and disadvantages of k-means algorithm according to Tayal and Raghuwanshi (2011). Firstly, the advantages: (1) k-means is a simple algorithm that has been adapted to many domains ma wrong. (2) k-means is more automated than making the threshold manually from an image or images. (3) This algorithm is a good candidate to be used as a continuation of the work relates to vectors that have the feature or vague (fuzzy) characteristics.

Next, the disadvantages are: (1) though it can be demonstrated that the procedure will always end, k-means clustering algorithm does not always find the most optimal configuration, which is related to the global objective function. (2) This algorithm is also very sensitive to randomly selected cluster centers at the beginning. K-means algorithm can be run several times to reduce the impact on this problem.

1. MSE = largenumber;
2. Select initial cluster centroids $\{m_j\}_j$
   K = 1;
3. Do
4. OldMSE = MSE;
5. MSE1 = 0;
6. For $j = 1$ to k
7. $mj = 0$; $nj = 0$;
8. Endfor
9. For $i = 1$ to n
10. For $j = 1$ to$k$
11. Compute squared Euclidean distance $d^2(x_i, m_j)$;
12. Endfor
13. Find the closest centroid m_j to x_i
14. $m_{j=} m_{j+} x_i$, $n_j = n_j + 1$;
15. MSE1 = MSE1 + $d^2(x_i, m_j)$;
16. Endfor
17. For $j = 1$ to k
18. $n_j = \max(n_j, 1)$; $m_j = m_j / n_j$;
19. Endfor
20. MSE = MSE1; while (MSE<OldMSE)

Figure 3 Traditional k-means Algorithm (Oyelade *et al.,* 2010)

METHODS

The method applied in this study generally includes three main stages: (1) data collection, (2) data pre-processing, and (3) data mining. First, data collected in this study was taken from Indonesia Stock Exchange Bursa Efek Indonesia website. Second, data pre-processing is the most important task in data mining. This stage is often said to take almost 80% of the total time or task in data mining. Techniques and methods to be applied in this stage must be precise and correct. Data pre-processing used in this study is based on the theory by Jiawei, Han and Michelin which includes: (1) Data Cleaning: filling in the missing values, repairing data errors, identify or remove outliers, and fixing inconsistent data. (2) Data Integration: merging related data from tables, databases, cube, or files. (3) Data Selection: selecting data only related to the process of analysis. The benefit of this step is to reduce less important or less relevant data in data mining processes. (4) Data Transformation: transforming data to support the process of analyzing the data that will be used.

Third, Data Mining is the primary stage of the entire task in this study. As with the data collection as well as data pre-processing, this stage also applies the theory by Jiawei Han and Michelin which include: (1) Data Mining, this stage is the stage for the implementation of the modeling used in data mining. In this study, the model applied is k-means cluster analysis. (2) Pattern Evaluation, this is an evaluation of the pattern that has been processed. (3) Knowledge Presentation, this is a presentation of the results of the data mining process.

RESULTS AND DISCUSSIONS

Application of cluster analysis in this study applied four clusters. Cluster analysis is implemented on two attributes, namely the volume and value of transactions of LQ45 stocks in the Indonesia Stock Exchange. Data source in this study were taken from the Indonesia Stock Exchange (LQ45, 2015). The data applied in this study were the data that were last updated on February 5, 2015.

The original data contained 27 attributes. In the study of cluster analysis, the software used is Rapid Miner studio. Pre-processing step will pick and choose three attributes for cluster analysis, namely (1) attribute code shares, (2) attributes the volume of transactions, and (3) the value attribute stock. Attribute of stock code will act as an identifier, while the volume attribute is an attribute that describes the number of shares traded and value is to describe the total value of transactions.

The second cluster analysis in this study implements similarities and dissimilarities between the measurements of data objects based on Euclidian distance measurement method. For example, i $= {}_{(i1}\chi, \chi_{i\,2,} \ldots, \chi_{i\,p)}$and j = $(\chi_{j\,1,} \chi_{j2,} \ldots, \chi_{j\,1)}$ are two of the objects described by attribute numerical p, then to measure the distance Euclidian between these objects is (Han & Kamber, 2012):

$$d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2}.$$

Similarity measurement technique using Euclidian's method as above also meet the mathematical properties such as the following (Han, 2012): (1) Non-negative: d (i, j) ≥ 0: Distance is not negative. (2) The identity of an indistinguishable: d (i, i) = 0: Distance of an object to itself is 0. (3) Symmetrical: d (i, j) = d (j, i): The distance is a function of symmetry. (4) Triangle inequality: d (i, j) ≤ d (i, k) + d (k, j): The distance the object i to j cannot be greater than the distance rotate through k.

The following plots below are the results of a cluster analysis study of LQ45 stocks in the Indonesia Stock Exchange on November 6, 2015 transaction.

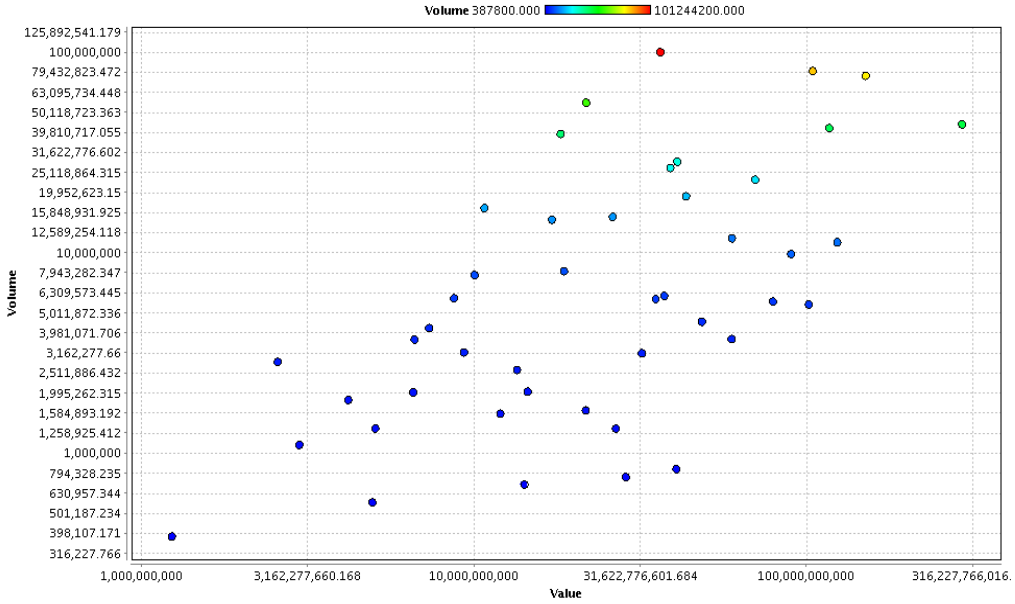

Figure 5 Plots Cluster Analysis Results of LQ45 Stocks in the Indonesia Stock Exchange for Transactions on November 6, 2015 Based on Volume

From the plot, it can be seen that the graph between volume and value, cluster 0 has 18 objects, Cluster 1 has 8 objects, cluster 2 has 11 objects, and cluster 3 has 8 objects. Cluster 0, looked dominant in terms of membership numbers objects compared with other clusters. Moreover, it looks very dominant in terms of the density of the distance between objects. From the findings of the two properties, then we can conclude that stocks in LQ45 that are the most sought after by investors are the combination of stocks with low value transactions and low volume transaction.

As with other studies, this study certainly is not perfect in the area of cluster analysis of LQ45stocks index in the Indonesia Stock Exchange. Some weakness potential in this study that is identified by the researchers is the need to get better cluster by comparing the accuracy of cluster analysis among various experiments by applying different number of clusters. Furthermore, to get better result, the study also needs to compare the accuracy with different cluster analysis, such as k-medoids, etc.

CONCLUSIONS

This cluster analysis could provide information more quickly and efficiently on the distribution map of LQ45 stocks in the Indonesia Stock Exchange. Results of cluster analysis LQ45 stocks in the Indonesia Stock Exchange provide information that is useful and quick visual to view a map of LQ45 stocks that soon became the target in decisions of stock investors.

REFERENCES

Athanasios, V., & Antonios, A. (2012). Stock market development and economic growth an empirical analysis. *American Journal of Economic and Business Administration, 4,* 135-143.

Gothai, E., & Balasubramanie, P. (2012). An efficient way for clustering using alternative decision tree. *American Journal of Applied Science, 9,* 531-534.

Han, J., & Kamber, M. (2012). *Data Mining: Concepts and Techniques* (4th ed.). San Francisco: Morgan Kaufmann Publishers.

Hossain, J., Fazlida Mohd Sani, N., Mustapha, A., & Affendey, L.S.(2013). Using feature selection as accuracy benchmarking in clinical data mining. *Journal of Computer Science, 9*,883-888.

IDX. (n.d.). Bagi Perusahaan. Retrieved from http://www.idx.co.id/Stocklist/LQ45/tabid/175/lang/en-US/language/en-US/Default.aspx

Kumar, S.P., & Ramaswami, K.S. (2011). Fuzzy modeled k-cluster quality mining of hidden knowledge for decision support. *Journal of Computer Science, 7,* 1652-1658.

LQ45. (2015, February 5). Retrieved from www.idx.co.id/id-id/beranda/publikasi/lq45.aspx

Oyelade, O. J., Oladipupo, O. O., & Obagbuwa, I. C. (2010). Application of k-Means Clustering algorithm for prediction of Students' Academic Performance. *International Journal of Computer Science and Information Security (IJCSIS), 7*(1). Retrieved on August 3, 2015 from http://arxiv.org/ftp/arxiv/papers/1002/1002.2425.pdf

Rui, X., Donald, W. (2005). Survey of Clustering Algorithms. *IEEE Transactions on Neural Networks*, *16*(3), 645 – 678.

Silwattananusarn, T., & Tuamsuk, K. (2012). Data Mining and Its Applications for Knowledge Management: A Literature Review from 2007 to 2012. *International Journal of Data Mining & Knowledge Management Process (IJDKP), 2*(5). Retrieved on August 3, 2015 fromhttp://arxiv.org/ftp/arxiv/papers/1210/1210.2872.pdf

Tajunisha, S. (2010). Performance analysis of k-means with different initialization methods for high dimensional data. *International Journal of Artificial Intelligence & Applications (IJAIA), 1*(4), 44-52. Retrieved on August 3, 2015 from https://www.academia.edu/12640770/Performance_analysis_of_k-means_with_different_initialization_methods_for_high_dimensional_data

Tayal, M.A., & Raghuwanshi, M.M. (2011). Review on Various Clustering Methods for the Image Data. *Journal of Emerging Trends in Computing and Information Sciences, 2 Special Issue.*

Wang, H., & Song, M. (December, 2011). Ckmeans.1d.dp: Optimal k-means Clustering in One Dimension by Dynamic Programming. *The R Journal, 3*(2), 29-32.