DATA MINING TECHNIQUE WITH CLUSTER ANAYSIS USE K-MEANS ALGORITHM FOR LQ45 INDEX ON INDONESIA STOCK EXCHANGE

A. Raharto Condrobimo[1,2,]
[1]Computer Science Department
BINUS Graduate Program - Doctor of Computer Science,
Bina Nusantara University,
Jakarta, Indonesia 11480
[2]Information System Department
School of Information System
Bina Nusantara University
Jakarta 11480, Indonesia
acondrobimo@binus.edu

Bahtiar Saleh Abbas
Computer Science Department
BINUS Graduate Program - Doctor of Computer Science,
Bina Nusantara University,
Jakarta, Indonesia 11480
bahtiars@binus.edu

Agung Trisetyarso
Computer Science Department
BINUS Graduate Program - Doctor of Computer Science,
Bina Nusantara University,
Jakarta, Indonesia 11480
atrisetyarso@binus.edu

Wayan Suparta
Civil Engineering Department
University of Technology Yogyakarta,
Yogyakarta Indonesia
drwaynesparta@gmail.com

Chul-Ho Kang
Department of Electronic and Communication Engineering
Kwangwoon University, South Korea
Chikang5136@kw.ac.kr

***Abstract*— This study aims to apply data mining techniques with cluster analysis on stock data registered in LQ45 in Indonesia Stock Exchange. The cluster analysis used in this method is k-means algorithm, the data in this research is taken from Indonesia Stock Exchange. The cluster analysis in this study analyzed the characteristics of data volumes and stock values, while the results in this study were presented in the form of cluster members visually. Therefore, this cluster analysis in this research can be used for quick and efficient identifier for each member of LQ45 index cluster based on share value for each cluster and its volume. The identification results can be used by beginner-level investors that begun to be interested in stock investments to help make informed decisions about stock trading on desired cluster groups.**

Keywords*— *data mining, cluster analysis, k-means, stocks

I. INTRODUCTION

Currently, it is common for professional stockbrokers to try to extract relationships from different stocks by analyzing past trading graphs thoroughly. In addition, more available stock system software predictions can be used by stock investors to help them generate fast stock market forecasts [1].

Stocks or shares, in this study we do not distinguish stocks and shares, is the relationship of ownership between the company and shareholders. There are two types of stock in the classification of shares in general, namely 1) preferred stock and 2) ordinary shares. Preferred stock is a stock that has a special right in the company (for example: distribution of previously received corporate profits rather than other shareholders) whereas ordinary shares are shares that have no more rights than the general right to obtain the profit in accordance with the profit-sharing schedule which held in the Annual General Meeting of Shareholders (AGMS). Compared to preferred shares with special interests that can be transferred to other parties for trading on the stock market, ordinary shares (hereinafter referred to as shares) have an advantage. Indonesia Stock Exchange (IDX) is the only stock market in Indonesia. IDX provides a mechanism for selling and buying shares for publicly listed companies listed on the Stock Exchange. Limited Liability Company (PT) is a legal entity to run a business consisting of share capital, which is part of its share owners. PT TBK is a company with limited liability company also public company status (Go Public).

In the capital market instruments traded, stocks are the main product. There are some derivatives that arise from transactions in the stock market. There are two ways in

978-1-5386-0954-5/18/$31.00 ©2018 IEEE

investing stocks, first, buying and store these shares so the benefit came from the profit distribution of dividends (dividends) and second, stock buying and selling back, the benefit will come from the deviation between the sale and purchase value (capital gain). Purchase of shares in general can be done in two ways, purchased when the stock will rise and start when Initial Public Offering (IPO) and purchased through the secondary market or through the stock market.

Shares are an investment only for the upper classes, this happens in the period before online. As technology grows, and the era of online trading is increasing where the transaction can use the Internet online network, stock transactions increasingly shifted into investment options for many people. This is supported by a minimum initial deposit more affordable for most people. Information about shares on the internet has become one of the sources of public information for who like to participate in stock investment, traders and investors. However, there still a debate about the updated information of the price or value of the stock over the internet [2].

Therefore, it makes sense to utilize data mining for stock market forecasting in mining historical data from the stock market to help determine better trading strategies. According to some researchers required procedures to seek comprehensive data mining, because of significant technical challenges and potential advantages, they are required to produce better forecasting results, in terms of accuracy, consistency, and reliability. Since the stock market index consists of many individual stocks and gives a broader picture of market movement than individual stock movements, the forecasting of the stock market index has attracted many researchers' interest.. [3].

The dynamics of financial markets play a major role in the functioning of the stock market, but the best predictive method has always been the topic of ongoing research and discussion. In the past decade, most methods have been based on stock index or index data, using unclear hybrid models or artificial neural networks to predict its future [4].

II. THEORITICAL BACKGROUND

A. Data Mining

Analyzing large amounts of data to efficiently extract important data as well as extracting hidden information from useful data by combining different techniques in different fields is the focus of data mining. Data mining techniques used include: pattern recognition, decision making, expert systems, knowledge discovery bases, artificial intelligence, and statistics. The main types of data mining include classification of mining, cluster mining, association rule mining, text mining, and image mining [3]. One of the data mining techniques such as clustering can be applied to uncover hidden knowledge of stock data. Clustering is the process of grouping a set of objects into a class of similar objects. Clusters are collections of data objects that resemble each other in the same cluster and are different from those in other clusters [5].

Data Exploration is a preliminary examination of the data to determine its main characteristics and determine the best approach for extracting meaningful information. The main purpose is to encourage in deciding the most appropriate pre-processing and data analysis techniques [6].

To overcome the pre-processing mistakes, there are several processes to be taken, such as cleaning, integration, transformation, reduction of news reports. This shows the missing value filling, combines the report by relevance and consolidates the data by replacing the original information using the news aggregator. Once the stored data is processed in pre-processing data stored in the data repository. The data repository contains data that has been cleared [7].

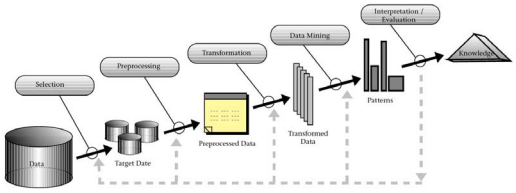

Figure 1. Data mining and knowledge discovery process of Database [11]

B. Cluster Analysis Or Clustering

The data source clustering is another data preprocess for multiple data mining sources. In contrast to the classification of data sources, this is a type of targeted learning. In other words, data clustering is a data cluster in accordance with the similarity between data without knowing the classes it has in advance [8] .

The results contribute to the existing literature by proposing a new time series model on groupings that (1) are more accurate than conventional approaches, (2) can be measured (on large datasets) because of the use of multi-resolution time at different levels of grouping, 3) can solve problems the limitations of comparative grouping algorithms in finding time series clusters that have a similarity of form. This important feature is very beneficial for the assessment of the cooperative in the stock market [9].

To manage and present complex datasets, Cluster analysis is often used to solve this problem [6]. To solve a problem around supervised learning or non-directed or unattended learning processes, cluster analysis can be considered the most popular technique. So, for every technique used in solving problems with such techniques, it will surely find a way to overcome the unlabeled data structures. [10]

Being one of the most prominent grouping techniques in science and technology, the k-means clustering algorithm will be used in this study. A collection of sorting algorithms will be used to group textual data collected from stock data on the Indonesia Stock Exchange [4].

A rich set of methods has been proposed in the past to group congruent data elements. The K-Means grouping (KM) is the most popular method that divides data into disaggregated groups using Euclidean distance between data elements and parallel cluster centers [10].

K-means grouping algorithm is used to optimize or improve the separation distance or similarity in the cluster. [11] The new average for each cluster will be calculated by this algorithm, by grouping the object into the cluster in the previous iteration. Then re-grouping is done for all objects using the new updated cluster center. The iteration process continues until it obtains a stable grouping, which means that the group formed from the last iteration is equal to the group formed in the previous iteration [11].

III. RESEARCH DESIGN AND METHOD

In this paper using four parts of cluster analysis applied in cluster analysis. Then implemented on two attributes, namely volume and transaction value on shares in the Liquid 45 or bluechip group in Indonesia Stock Exchange. The data used was taken form Indonesia Stock Exchange (http://www.idx.co.id/id-id/beranda/publikasi/lq45.aspx), last updated on November 5, 2017.

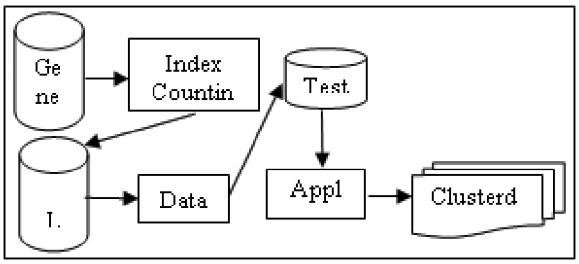

Figure 2. The Purpose Model

Figure 4 shows the proposed data mining model that applies K-means clustering algorithm. The data base on general stock list consist of 500 list. We use selected dataset in LQ45 list with two parameter which is volume and value.

IV. RESULT AND DISCUSSION

In this cluster analysis paper with K-Means, using software to perform the mining process is Rapid miner studio. At the pre-processing stage determine the attributes for cluster analysis, i.e. 1) the code of the stock attribute, 2) the transaction volume attribute, and 3) the attribute of the value of the stock. The attribute used as the identifier is the attribute of the stock code, while the volume attribute describes the number of shares traded and the attribute value of the stock represents the total value of the transaction.

In this research the second cluster analysis is used Euclidian distance measurement method, by applying equations and inequalities between measurement data objects. So for example, i = (i1 χ, χ i 2, ..., χ ip) and j = (χ j 1, χ j2, ..., χ j 1) are two of the objects described by the numerical attribute p, then to measure the Euclidian distance between these objects is [11]:

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

.

The Euclidean method used as a similar measurement technique as above also satisfies the mathematical properties, as follows: [11]:

- Positive: d (i, j) ≥ 0: Distance is must positive.
- An indistinguishable identity: d (i, i) = 0: Distance object to itself is 0.
- Symmetrical: d (i, j) = d (j, i): The distance is the symmetry function.
- Triangle inequality: d (i, j) ≤ d (i, k) + d (k, j): The distance of object i to j cannot be greater than the distance played through k

The results of cluster analysis of 45 blue chip stocks in Indonesia Stock Exchange on November 6, 2017 transactions are as follows.

The Euclidean n method used as a similar measurement technique as above also satisfies the mathematical properties, as follows: [11]:

- Positive: d (i, j) ≥ 0: Distance is must positive.
- An indistinguishable identity: d (i, i) = 0: Distance object to itself is h 0.
- Symmetrical: d (i, j) = d (j, i): The distance is a function of symmetry.
- Triangle inequality: d (i, j) ≤ d (i, k) + d (k, j): The distance of object i to j cannot be greater than the distance played through k

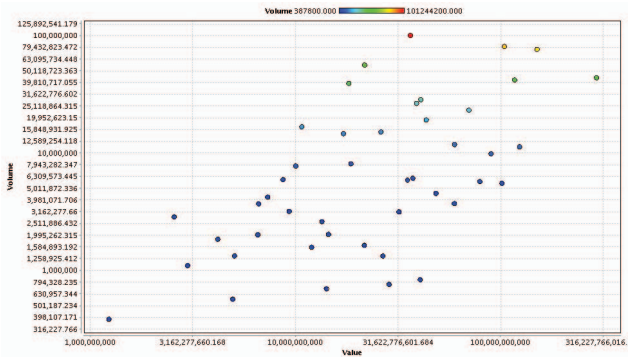

Figure 5. Plots cluster analysis results of 45 blue chip stocks

The graphic between volume and value is visible in the plot, 18 objects in Cluster 0, 8 objects in Cluster 1, 11 objects in Cluster 2, and 8 objects in Cluster 3. Cluster 0, looks dominant in terms of number of object membership compared to other clusters. In addition, the density of the distance between objects also looks very dominant. The conclusion drawn from the findings of both traits is that the stock in LQ 45 is the most desirable investor is a combination of stocks with low value transactions and low transaction volumes. Like many other studies, this study is still far from being a perfect study in conducting cluster analysis of 45 blue chip stocks on the Indonesia Stock Exchange. Some of the more developed potentials in this study, identified by researchers, is the need to compare the accuracy of cluster analysis when research is done by comparing experiments using different algorithms. Another potential weakness is the

need for comparative studies using other cluster analyzes, such as k-medoids and others.

V. CONCLUSION

Using cluster analysis in this study result with the ability to provide information quickly and efficiently for potential novice investors on the distribution map of Liquid 45 shares or bluechip stocks in Indonesia Stock Exchange.

The cluster analysis of 45 blue chip stocks in the Indonesia Stock Exchange provides useful and quick information visually to see the map of 45 blue chip stocks divided into four parts according to the needs in stock price attributes and share transaction value so as to provide information quickly and accurate to quickly become the target of stock investors' decisions.

REFERENCES

[1] Y. Luo, J. Hu, X. Wei, D. Fang, and H. Shao, "Stock trends prediction based on hypergraph modeling clustering algorithm," in *2014 IEEE International Conference on Progress in Informatics and Computing*, 2014, pp. 27–31.

[2] H. Leung and T. Ton, "The impact of internet stock message boards on cross-sectional returns of small-capitalization stocks," *J. Bank. Financ.*, vol. 55, no. December 1997, pp. 37–55, 2015.

[3] X. Zhong and D. Enke, "A comprehensive cluster and classification mining procedure for daily stock market return forecasting," *Neurocomputing*, vol. 267, pp. 152–168, Dec. 2017.

[4] E. N. Desokey, A. Badr, and A. F. Hegazy, "Enhancing stock prediction clustering using K-means with genetic algorithm," in *2017 13th International Computer Engineering Conference (ICENCO)*, 2017, pp. 256–261.

[5] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, "Analyzing undergraduate students' performance using educational data mining," *Comput. Educ.*, vol. 113, pp. 177–194, Oct. 2017.

[6] M. S. Packianather, A. Davies, S. Harraden, S. Soman, and J. White, "Data Mining Techniques Applied to a Manufacturing SME," *Procedia CIRP*, vol. 62, pp. 123–128, 2017.

[7] R. Mythily, A. Banu, and S. Raghunathan, "Clustering Models for Data Stream Mining," *Procedia Comput. Sci.*, vol. 46, no. Icict 2014, pp. 619–626, 2015.

[8] R. Wang *et al.*, "Review on mining data from multiple data sources," *Pattern Recognit. Lett.*, vol. 0, pp. 1–9, Jan. 2018.

[9] S. Aghabozorgi and Y. W. Teh, "Stock market co-movement assessment using a three-phase clustering method," *Expert Syst. Appl.*, vol. 41, no. 4 PART 1, pp. 1301–1314, 2014.

[10] V. Vijay, V. P. Raghunath, A. Singh, and S. N. Omkar, "Variance Based Moving K-Means Algorithm," in *2017 IEEE 7th International Advance Computing Conference (IACC)*, 2017, no. i, pp. 841–847.

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