Role of clustering based on density to detect patterns of stock trading deviation

Alvida Mustika Rukmi¹, Soetrisno², Abirohman Wahid³
¹,²,³Mathematics Department, Institut Teknologi Sepuluh Nopember (ITS) Indonesia
Campus ITS, Sukolilo, Surabaya, 60111, Indonesia
alvidamr@gmail.com¹, soetrisno@matematika.its.ac.id², wabirohman@gmail.com³

Abstract. The pattern of deviation patterns can be identified from the results of cluster transactions and transactions that are transaction irregularities, will be detected. DBSCAN as a density-based clustering algorithm forms clusters that agglomerate and make it easier to detect unclustered data, which is considered as data noise (data outlier). The nature of density in the data clamping process will make it easier to determine noise data objects. The DBSCAN has two parameters, Eps and MinPts. The values entered in both parameters play a role in forming clusters. Stock trading transactions are stated as data objects to be clustered. The noise from clustering with DBSCAN shows outlier transactions, which have different patterns with ordinary transactions. In the results of this clustering, the stock transaction pattern which includes outliers is obtained, marking the close occurs. This result can help to detect stock price manipulation in outlier transactions carried out by securities brokers.

1. Introduction

Based on the Capital Market Law No. 8 of 1995 (Capital Market Law) Capital Market crime can be divided into 3 categories, namely Fraud, Market Manipulation and Insider Trading. Some types that can be categorized as market manipulation crime versions of the Capital Market Law are creating a false capital market picture such as 1) Conducting securities transactions that do not result in changes in ownership, or 2) Selling offers or buying securities at certain prices, while other parties who are conspirators also make a buy offer or sell offer at the same price (vide Article 91 UUPM). Other activities are to carry out two or more securities transactions on the Stock Exchange, causing a fixed, up or down effect price, with the aim that the other party is affected to buy, sell or hold the effect. As a result, the price of the securities is not based on actual buying or selling requests. (vide: Article 92 of the Capital Market Law) [1].

Fraud on stock transactions usually occurs when securities brokers or investment managers indicate to manipulate their customers into stock trading regardless of the customer's own interests. If capital market players are not protected from people who are able to manipulate share prices, it can influence the people of capital market players. The amount of fraud may not be too much but if there is fraud, it will cause a large loss. Stock transaction data can be used to identify stock price manipulation through data mining. Data mining can be done with the techniques found in data mining. Density Based Spatial Clustering of Application with Noise (DBSCAN) which has a density-based clustering. Every transaction that is conducted by clustering is possible there is a data object in this case the transaction, not classified into a cluster. Every transaction that affects prices is usually carried out with a value and time that is not reasonable compared to transactions in general can cause suspicion for observers of stock trading transactions.

In addition, transactions are also observed which correspond to the definition of manipulation of manipulation activities as follows.

1. Painting the tape is the trading activity between one securities account and another securities account that is still in the possession of one party or has an attachment in such a way as to create a false trade.
2. Pools (pump-pump manipulation) is a large fund association by a group of investors where the fund is managed by an intermediary or someone who understands market conditions.
3. Marking the close is to engineer the request or offer price at the time of or nearing the close of trading with the aim of forming a price of securities or a higher opening price on the trading next day.

2. Discussion

2.1 Preprocessing Data
This stage is carried out to explore the data that will be used. The object data used is LQ45 data, in the form of transaction data on data within 20 working days of the exchange. The list of shares included in the LQ45 list has been made by the Indonesia Stock Exchange, which is a stock with a level of efficiency and price manipulation. Data sources are available in the HOTS application for customers. The data used is transaction data where the attributes needed for each data record are transaction time, transaction date, transaction volume, buyer and seller. Transaction time and transaction date are used to obtain data noise, use unit of measurement in seconds. It can improve accuracy when clustering process. The transaction value is the amount used to buy and sell shares. Buyers are people who intend to have money with money they have. While the seller is the person who provides the shares he has. Transactions between buyers and sellers occur in price deal between sellers and buyers. The process use many attributes on the stock transaction to make clustering.

2.2 DBSCAN Algorithm
After going through pre-processing, data clustering is done with the aim of getting noise. Noise in the sale and purchase transaction data is considered an unusual transaction. Data noise results from DBSCAN cluster results in securities transaction data can be defined as transactions that have a large difference compared to all other transactions so that transactions like this are unusual transactions.

The following are the algorithms used in the program to get transactions that are considered suspicious. The following algorithm becomes the main class in the program. After inputting the data, the DBSCAN cluster is performed with the algorithm described after the main algorithm. The result of DBSCAN clustering is the transaction object which is then collected in table form, this transaction has the value and frequency of occurrence as the respective noise.

DBSCAN algorithm requires input transaction data with the attributes previously described, Radius parameter $\varepsilon$ as a reference for cluster formation and $\text{minPts}$ as the minimum object reference in cluster formation. The value of $\varepsilon$ has its own unit which considers the comparison between the number of transactions and the time of the transaction described after the DBSCAN algorithm. Object $p$ is the core point for the formation of each cluster collected in $C$, then $p$ 'is a p environment which then becomes the core point for the formation of the next cluster collected in the N clusters.

Processing to make clusters using DBSCAN algorithm
Input:
- a. $D$: Data set number n objects.(share data per day / month, the object in question is a transaction)
- b. $\varepsilon$: radius parameter(has its own unit, adjusted to the time and value of the transaction)
- c. $\text{minPts}$: minimum number of objects

Output:
Set data with the DBSCAN cluster

```
Algorithm DBSCAN (D, $\varepsilon$, $\text{minPts}$)
//all objects $O_c$ in D as a unvisited point.
Begin
For all objects $O_c$ in D do
If select randomly unvisited point as $O_c$ in D
```
Invoke `cluster_making` to construct a cluster

**Method** `cluster_making ( \( O_c \), D, \( \varepsilon \), minPts )`

**Begin**

Mark \( O_c \) as visited point

**For every** \( O_i \)

\[
\text{If } \varepsilon \geq \sqrt{(O_{i,t} - O_{c,t})^2 + (O_{i,v} - O_{c,v})^2} \\
\text{Mark } O_i \text{ in } \varepsilon \text{–neighborhood of } O_c, N_{\varepsilon} (O_c) \\
|N_{\varepsilon} (O_c)| = |N_{\varepsilon} (O_c)| + 1 \\
\text{If } |N_{\varepsilon} (O_c)| \geq \text{minPts} \\
\text{Create a new cluster_id, C, and insert } O_c \text{ to C}
\]

**For every** \( O_i \) in \( D' \) is object set in \( N_{\varepsilon} (O_c) \)

\[
\text{If } O_i \text{ is unvisited point} \\
\text{Mark } O_i \text{ as visited point} \\
\]

**For every** \( O_i \) in \( D' \)

\[
\text{If } \varepsilon \geq \sqrt{(O_{i,t} - O_{c,t})^2 + (O_{i,v} - O_{c,v})^2} \\
\text{Mark } O_i \text{ in } \varepsilon \text{–neighborhood of } p, N_{\varepsilon} (p) \\
|N_{\varepsilon} (O_i)| = |N_{\varepsilon} (O_i)| + 1 \\
\text{If } |N_{\varepsilon} (O_i)| \geq \text{minPts} \\
\text{Insert object to } D' \\
\]

\[
\text{If } O_i \text{ is not object in } C \\
\text{insert } O_i \text{ to } C
\]

**End For**

Output C

**Else** as outlier point

**Until** all object as visited points

**Annotation** :

\( O_i = i\text{-th Object} \)
\( O_{c,t} = \text{Core object or point on time attribute} \)
\( O_{i,t} = i\text{-th object in time attribute} \)
\( O_{c,v} = \text{Core object or point on value attribute} \)
\( O_{i,v} = i\text{-th object on value attribute} \)

To get a comparison of transactions, the values of \( O_{i,v} \) and \( O_{c,v} \) are divided by the results of the largest value division with total time. Then obtained the same comparison between transaction time and volume.

DBSCAN is a density-based clustering algorithm because this algorithm searches a number of clusters starting from estimating density on objects that are related to each other [16]. So that if there is an increase in the number and value of stock transactions in a certain period of time, then the transaction will be classified into clusters. For high value transactions but in fair conditions such as when the financial statements of a company with high profit are published, the publication of a company's performance news, and capital market experts advise capital market players, these transactions will be classified into a cluster because the number of transactions is sold at a certain value and time period. Conversely, if there is an unnatural transaction in this case, the value is large and there is no transaction density, the transaction will be classified as an outlier.

After going through pre-processing, data clustering is done with the aim of getting outliers. Outliers in the sale and purchase transaction data are considered as transactions that do not have similarities with transactions generally.
2.3 Clustering using DBSCAN algorithm
Transaction data that recorded as Company-1 share transaction data has transactions 44,769 transactions from May 8, 2018 to June 8, 2018. The stock price at the opening of trading on May 2, 2018 is Rp. 1,160.00 while at the closing on June 8 2018 is Rp. 1,145.00 so that in the time span of this data there is a price decrease of Rp. 15.00

![Figure 1. Company-1 stock data](image)

Table 1. Input Data

| DateTime         | Price | Volume | Buyer | Seller | Buyer Type | Seller Type |
|------------------|-------|--------|-------|--------|------------|-------------|
| 07/05/2018 09:00:06 | 1150  | 5      | PD    | PD     | D          | D           |
| 07/05/2018 09:00:09 | 1140  | 37     | KK    | YP     | D          | D           |
| 07/05/2018 09:00:09 | 1145  | 1      | PD    | YP     | D          | D           |
| 07/05/2018 09:00:09 | 1145  | 1      | DH    | YP     | D          | D           |
| 07/05/2018 09:00:09 | 1145  | 1      | DR    | YP     | D          | D           |
| 15/05/2018 09:01:48 | 1135  | 44     | YP    | YP     | D          | D           |
| 15/05/2018 09:01:52 | 1140  | 9      | NI    | YP     | D          | D           |
| 15/05/2018 09:01:52 | 1135  | 1      | NI    | CG     | F          | D           |
| 15/05/2018 09:01:52 | 1135  | 1      | NI    | BK     | F          | D           |
| 15/05/2018 09:01:52 | 1135  | 4      | NI    | YP     | D          | D           |
After experimenting several values of \( \epsilon \) and minPts, the following graphs and tables are obtained which have the value of the number of outlier transactions that appear.

![Figure 2. Number of outliers for some \( \epsilon \) and minPts values in Company-1 shares](image)

Result of clustering using DBSCAN algorithm, give number of outliers for some values on two parameters, \( \epsilon = 10000, 25000, 5000 \) and minPts = 10, 25, 50, 100. The result shown on Table 2.

**Table 2. Number of outliers for some \( \epsilon \) and minPts \(^{'}\) s values in Company-1**

| \( \epsilon \) | minPts | 10  | 25  | 50  | 100 |
|----------------|--------|-----|-----|-----|-----|
| 10000          |        | 1051| 2064| 3221| 4297|
| 25000          |        | 307 | 656 | 1144| 1605|
| 50000          |        | 167 | 372 | 514 | 745 |

Data is processed with the value \( \epsilon = 50000 \) with minPts = 25 and the value \( \epsilon = 50000 \) with minPts = 10. After the program is run, the outliers are obtained.

![Figure 3. Results of clustering for a value of \( \epsilon = 50000 \) with minPts = 25](image)
Cluster results were observed on May 30 and May 31. On May 30, transactions in the range of 1,000,000 to 1,500,000 became outlier transactions because these transactions were not dense enough to form a cluster, whereas on May 31 at this price range the transaction is classified into the 17th cluster because there are transactions that are sufficiently solid to form a cluster.

This density occurred to coincide with the publication of the announcement regarding dividend distribution by the company. After holding a general meeting of shareholders. Transaction density in these conditions is considered reasonable in line with increasing buying interest in this stock.
Table 2 shows the results of the securities broker with outlier transactions. The intermediary effect is sorted by the frequency of the appearance of the number of transactions in the outlier. After that, the securities broker is observed with the outlier transaction with the most occurrence and the largest transaction value both as a buyer and seller.

In this research, Company-1 shares were observed with two different parameter values so that the outlier transaction conditions were obtained as follows.

Table 3. Accumulation of transaction pattern

| Parameter | Securities Broker (buyer) | Price Increase | Securities Broker (seller) | Price Decline | Marking the close |
|-----------|---------------------------|----------------|----------------------------|---------------|------------------|
| $\varepsilon = 50000$ with $\text{minPts} = 25$ | AG | 12/23 | 3/15 | | |
| | DH | 11/13 | 11/11 | PD | 6/26 |
| | | | | CC | 14/31 |
| | TOTAL | 23/36 | 20/57 | 35/64 |
| $\varepsilon = 50000$ with $\text{minPts} = 10$ | AG | 8/15 | 1/8 | | |
| | DH | 9/11 | 1/7 | DB | 3/19 |
| | | | | GR | 9/15 |
| | TOTAL | 17/26 | 12/34 | 16/35 |

On the Table 3 show outlier transactions tend not to result in changes in the expected increase by buyers or sellers on the next day. For buying transactions which are outliers there is not much increase in the next trading day, so also the sales outliers do not often decrease in prices on the following trading day.

Based on the two parameter values, the securities broker as the buyer with the biggest and most outlier transactions is AG and DH. Outlier transactions carried out by securities brokers with the highest number of outlier transactions or the biggest outlier transaction values both as buyers and sellers tend to increase at the opening of the next trading day. Of the total 99 transactions conducted by securities brokers suspected of marking the close, there were 51 transactions in accordance with the definition of marking the close.

3. Conclusion

The results of clustering with the DBSCAN algorithm on stock trading (capital market) transactions conducted by securities brokers produce outlier transactions. The result look for outlier transactions with 2 parameters in Company-1 shares, is conducted.

In Company-1 shares, the parameter value $\varepsilon$ is 50,000 and the minPts value is 25. After clustering with the two parameters, 24 clusters are obtained. Then 372 transactions are classified as outliers from a total of 44,769 transactions. There obtained 35 transactions from 64 transactions according to the definition of marking the close.

Then the parameter value $\varepsilon$ is equal to 50,000 and the minPts value is 10. After clustering with the two parameters, a cluster of 23. Then we get 167 transactions classified as outliers out of a total of 44,769 transactions. There are 16 transactions obtained from 35 transactions according to the definition of marking the close.
References

[1] Undang-Undang No.8 Tahun 1995 Tentang Pasar Modal.
[2] Han J., Micheline Kamber, Jian Pei. 2012. Data Mining: Concept and Techniques, Third Edition. Waltham: Morgan Kaufmann Publishers.
[3] Golmohammadi Koosha, Osmar R. Zaiane. 2012. Data Mining Applications for Fraud Detection in Securities Market. Edmonton. Department of Computing Science University of Alberta, Canada.
[4] Golmohammadi Koosha, Osmar R. Zaiane. 2014. Detecting Stock Market Manipulation using Supervised Learning Algorithm. Edmonton. Department of Computing Science University of Alberta, Canada.
[5] Diaz David, Babis Theodoulidis, Pedro Sampaio. 2011. Analysis of Stock Market using Knowledge Discovery Techniques Applied to Intraday Trade Prices. Manchester. Manchester Business School The University of Manchester.
[6] Ferdousi Zakia, Akira Maeda. 2006. Unsupervised Outlier Detection in Time Series Data. Kusatsu. Department of Media Technology Ritsumeikan University, Japan.
[7] Ferdousi Zakia, Akira Maeda. 2006. Mining Outliers for Fraud Detection in Financial Data. Kusatsu. Department of Media Technology Ritsumeikan University, Japan.
[8] Aguinis Herman, Gottfredson R.K., Joo Harry. 20013. Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. Department of Management and Entrepreneurship, Kelley School of Business, Indiana University, USA.
[9] PT. Adhi Karya (Persero) Tbk.: Laporan tahunan 2016.
[10] PT. Perusahaan Perkebunan London Sumatra Indonesia Tbk.: Laporan Tahunan 2014
[11] PT. Sri Rejeki Isman Tbk.: Laporan Tahunan 2015.
[12] Fahim A.M., Salem A.M., Torkey F.A., Ramadan M.A. 2006. Density Clustering Based on Radius of Data. World Academy of Science, Engineering and Technology 22.
[13] Cai Fan, Le-Khae Nhien-An, Kechadi M-Tahar. 2015. Clustering Approaches for Financial Data Analysis. School of Computer Science & Informatics, University College Dublin, Ireland.
[14] Aitken M.J., Harris F.H.deB., Ji Shan. 2009. Trade-Based Manipulation and Market Efficiency: A Cross-Market Comparison. Australian School of Business.
[15] Setiyono, Budi danImam Mukhlash. 2005. Kajian Algoritma GDBScan, Clarans dan Cure untuk Spatial Clustering. Limits. Vol.2, No.2. Institut Teknologi Sepuluh Nopember. Surabaya.
[16] Suganthi R., P. Kamalakannan. 2015. Analyzing Stock Market Data Using Clustering Algorithm. International Journal of Future Computer and Communication, Vol. 4, No.2.