HASIL CEK_ICAIBDA 21 Paper

by Icaibda 21 Paper

Submission date: 16-Oct-2021 09:56AM (UTC+0700)

Submission ID: 1675226512

File name: ICAIBDA 21 Paper [Camera Ready] (Endang dkk).docx (3.21M)

Word count: 3310

Character count: 17591

Outlier Detection Using K-Means Clustering with Minkowski-Chebyshev distances for Inquiry-Based Learning Results in Students Dataset

1st Endang Wahyuni
Master of Mathematics Education
Ahmad Dahlan University
Yogyakarta, INDONESIA
endang1907050001@webmail.uad.ac.id

2nd Sugiyarto Surono
Dept. Mathematic FAST
Ahmad Dahlan University
Yogyakarta, INDONESIA
sugiyarto@math.uad.ac.id

3rd Joko Eliyanto
Master of Mathematics Education
Ahmad Dahlan University
Yogyakarta, INDONESIA
joko1907050003@webmail.uad.ac.id

Abstract*— Outlier appears as an extreme value but often contains very important information, so it is necessary to be studied whether the data remains used or issued. Outlier detection is a hot topic for the study. Increasing new technologies and various applications cause increased requirements of outlier detection. The Outlier method is successfully applied in various fields, namely: economy, business, health, space, geology, and education. Implementation of an outlier of analysis in the education field is often applied to the evaluation of the learning model. Inquire -based learning model is an important component in education renewal. Learning by this method encourages learners to learn mostly through their active involvement. This study aims to discuss outlier detection by using the K-Means Clustering method on the inquiry-based learning results in students. This study detects outliers with the K-means method using Minkowski-Chebyshev distance. The result of the proposed method will be compared with the extremes of standard deviation (ESD), Box-Plot, and K-Means Clustering using Euclidean distance. The outlier detection results using K-Means Clustering with the Minkowski-Chebyshev and Euclidean distance produce the same result that can detect 3 data as an outlier that is the student with the ID number 7 Exam Value 7.5, ID number 42 Exam Value 9.0, and ID number 72 with the value of 13.5. While the ESD method and Box-Plot are unable to detect any outlier.

***Keywords* — Outlier Detection; K-Means Clustering; Inquiry based learning; ESD; Box-Plot;**

I. Introduction

Data that have large volumes, various types of data, and very fast data speeds called Big Data[1] [2] [3]. Variable data can be represented as variables, while variables are seen as dimensions [2] clustering is one of the unsupervised learning methods, wherein terms of data cluster or value do not have a target or have no class label [4]. Clustering is the process of grouping the data into groups or clusters. Each cluster has data that has a high resemblance, and between clusters has a low resemblance [4].

The Outlier Detection method is the most important method of data analyzing, such as decision making, grouping, and pattern classification. Outlier is defined as observations that do not match the overall pattern of grouping [1] [2] [3] [5] [6]. Outlier detection is an important subject in data mining. Outlier detection is widely used to identify and eliminate ordinal or irrelevant objects of data set [2] [6] [7] [8][9]. However, the main challenge of outlier detection is increased complexity due to diverse datasets and dataset size [7]. The outlier factor of the cluster determines the degree of difference from a cluster of the whole dataset [10]. Implementation of outlier analysis in the education field is often applied to the evaluation of the learning model. Inquire-based learning(IBL) model is an important component in educational renewal. The Inquiry learning model is a process of learning-oriented to the activity of the learners in the process of investigation and discovery of solutions from the issue issued [15]. Inquiry learning puts learners as a subject of learning. Learners play a role to discover the core of the material, while educators play a participant as actors and act as facilitators, motivators for learners. The IBL model stage consists of three stages: producing hypothesis and investigations and discovery and reflection. Based on the relevance of each Inquiry-based Learning step with the indicator of critical thinking ability of student mathematics. In addition, Inquiry Learning focuses on students inactivating in an investigation that leads to students to logical generalizations. These activities have the potential in facilitating the enhanced critical thinking ability. Learning models that give students more freedom have the potential to produce very varied results. Including the occurrence of outliers. This study attempts to detect outliers in the inquiry-based learning dataset. The existing clustering method has fixed outlier determination criteria. This often makes data that is actually quite different from other data not detected as outliers. A method that has a flexibility threshold is needed to determine outliers. The K-means clustering method was chosen to be developed as an outlier detection because of its ability to group data based on the similarity between the data. The greater the difference in a dataset member, it can be concluded that the data is an outlier. The value constraint that determined whether a dataset member was an outlier or not is called a threshold. In the k-means method, the threshold setting is very possible and depends on the distance function used. This study also tried to experiment with the development of an outlier detection method on K-means clustering using Minkowski-Chebyshev distances. This is one element of the novelty offered in this research compared to other studies. The results of the proposed method are then compared with the commonly used outlier detection methods, ESD, Box-Plot rule, and K-Means Clustering with Euclidean distance.

II. METHODS

A. Outlier Detection

Outlier detection is obtained data that appears with extreme values both univariate and multivariate. The extreme is a far or different value at all with most other grades in its group [2]. Outlier detection is an important subject in data mining. Outlier detection is widely used to identify and eliminate ordinal or irrelevant objects of data set [2] [6] [7] [8][9]. However, the main challenge of outlier detection is when increased complexity due to diversity dataset and dataset size [7]. The outlier factor of the cluster determines the degree of difference from a cluster of the whole dataset [10]. Outlier detection is used to find fraudulent data. Researchers research data groupings and Outlier Detection process [6][10]

B. K-Means Clustering

The K-Means clustering the data that exist into some clusters with the criteria in the same cluster has the same characteristics, and have different clusters with data in other clusters. This algorithm is most widely used among all clustering algorithms due to its efficiency and its simplicity [8]. In this algorithm, the number of clusters is assumed to remain.

Suppose D is a dataset with a number of n rows, eg $C_1, C_2, \cdots, C_m$ a is a separate cluster inside D. Then the error function is defined as follows:

$$E = \sum_{i=1}^{m} \sum_{x \in C_i} d(\vec{x}, \mu(C_i)) \qquad (1)$$

$\mu(C_i)$ = cluster center (centroid) of C_i

$d(\vec{x}, \mu(C_i))$ = the distance between data $\vec{x}$ to the centroid of cluster $\mu(C_i)$

- Euclidean Distance

The Euclidean Distance is one of the distance calculation methods used to measure the distance of 2 (two) fruit points in Euclidean Space (covering the two-dimensional euclidean field, three dimensions, or even more). To measure the level of similarity of the data with the Euclidean Distance formula used the following formula:

$$d(\vec{x}, y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2} \qquad (2)$$

where $x = (x_1, \cdots, x_m)$ and $y = (y_1, \cdots, y_m)$ are two input vectors with quantitative features m. In the Euclidean distance function, all features contribute the same at the function value [11] [12].

- Minkowski-Chebyshev Distance

Rodrigus (2018) raises a new distance that is a combination of Minkowski and Chebyshev. The combination of Minkowski and Chebyshev distance is shown in the following definitions:

$$d(w_1, w_2, r) = w_1 \left(\sum_{i=1}^{m} |x_i - y_i|^r \right)^{\frac{1}{r}} + w_2 (max_{i=1}^{m} |x_i - y_i|), 1 \leq k \leq m \qquad (3)$$

where $x = (x_1, \cdots, x_m)$ and $y = (y_1, \cdots, y_m)$ are two input vectors with quantitative features m.

C. Box-Plot-Rule

The standard Box-Plot, each has a top(U) and bottom(L) limit, defined as:

$$U = Q_3 + 1.5(Q_3 - Q_1)$$
$$L = Q_1 - 1.5(Q_3 - Q_1) \qquad (4)$$

the value that falls outside the limit is considered as outliers. This rule has a higher chance to detect false outliers than typical informal test [13]. Here Details How to Specify the Restrictions that can be seen in Figure 1.

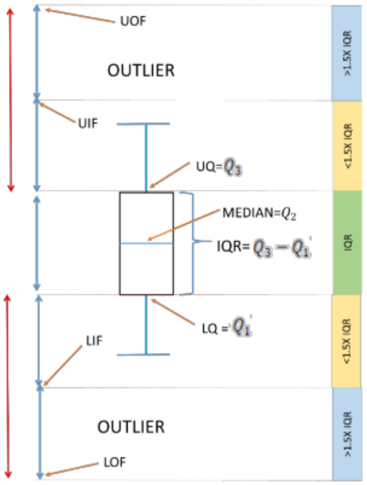

Figure 1. Determining the Boundaries on The Box-Plot-Rule Rules

D. ESD Test

Standard Deviation is a test based on the extreme studentized deviation (ESD) which is quite good at detecting outliers in a random normal sample. Defined x_j as outlier when:

$$G_j = \max_i \left\{ \frac{|x_i - \bar{x}|}{s_j} \right\}, i = 1, \cdots, m \qquad (5)$$

x_j is declared as outlier when $\frac{|x_i - \bar{x}|}{s_j}$ is largest. In principle, if G_j doesn't exceed the critical value, then no x_j need not be selected. If this test finds outliers, further outlier testing is carried out by deleting observations x_j and repeating the process on the remaining n-1 observations [14].

E. Inquiry -based learning

Inquire-based learning is oriented to the activity of the learners in the process of investigation and discovery of solutions from the issue issued [15]. According to [16] Learning outcomes is an overview of the ability of learners in meeting the stage of learning achievement in a basic competence (KD). Learning outcomes can be used as a benchmark or criteria in achieving educational goals. Inquire learning puts learners as a subject of learning. Learners play a role to discover the core of the material, while educators play a participant as actors and acts as facilitators, motivators for learners. Inquiry-based learning is an important component in educational renewal. Because learning with this method encourages learners to learn mostly through their active involvement, learners are expected to take the initiative, they are trained how to solve the problem, make decisions, and acquire skills. Educators encourage learners to have real and creative experiences.

Inquire-based learning is an important component of educational reform. Because learning whit this method encourages students to learn mostly through their active involvement, students are expected to take the initiative, they are trained how to solve problems, make decisions, and acquire skills/ educators encourage students to have real and creative experiences.Standard Deviation is a test based on the extreme studentized

III. RESULT AND DISCUSSION

In this study, the type of data used is secondary data obtained from the official page https://archive.ics.uci.edu/ml/index.php. The data used are the learning outcomes of undergraduate students using the Deeds e-learning (Digital Electronics Education and Design Suite) with an inquiry-based learning model (IBL) [17]. Deeds is a simulation application for e-learning in digital applications. This application provides learning materials through a special browser for students, and students are asked to solve various problems with different levels of difficulty. This application has been effective, for more than ten years, in teaching and improving student learning outcomes because it provides a highly interactive simulator. Digital electronics courses are organized in separate theory and laboratory sessions where students work with Deeds simulators. The inquire-based student learning process is applied in the context of education with the Deeds simulator.

At the beginning of each session, problem-solving exercises are given to students. For each exercise, students follow a learning process that involves understanding a given problem and dividing it into various tasks, making observations in a simulated environment, conducting experiments to find the answers, and finally explaining and justifying their solutions and the methods used. The final exam questions discuss the concept of each practice session. The exam questions consist of 6 practice sessions with a total of 16 description questions with each item having a different value according to the weight of the question.

At this stage is the process of selecting the relevant data. Irrelevant data will be removed. The data is selected and selected according to the calculation, where irrelevant data will be removed from the study, so it must be selected first which data will be used and which will be removed from the study. Many students do not take the exam, so the student data will be deleted from the exam results. Of the 115 students who attended the training, 22 students did not take the exam and only 93 students who took the final exam. So that 22 students who do not take the final exam will be removed or deleted. The results of student selection are in table 1.

Tabel 1. The Results of Student Selection

Student ID	Score	Student ID	Score	Student ID	Score
1	94,5	37	30,0	72	95,0
2	44,0	38	41,5	73	49,0
3	85,0	39	83,5	74	18,0
4	30,0	41	98,0	75	87,5
5	38,5	42	22,5	76	93,5
6	82,0	44	97,5	77	66,5
7	78,0	45	46,0	78	78,0
8	8,5	46	22,0	79	84,5
9	18,5	47	28,0	80	51,5
10	59,0	48	71,5	81	87,0
11	60,0	49	30,5	82	13,5
12	40,5	51	9,0	83	52,0
13	90,0	52	36,0	85	82,0
14	64,0	53	70,5	86	94,0
15	67,5	54	39,0	87	74,5
16	67,0	55	36,5	88	96,0
17	97,0	56	84,0	89	16,5
18	62,0	57	23,0	91	66,5
19	50,0	58	48,0	92	35,0
20	97,5	59	40,0	93	66,0
22	40,0	60	16,5	94	92,5
24	70,5	61	57,5	95	52,5
25	57,5	62	43,0	96	77,5
27	74,5	63	69,0	98	66,5
28	79,5	64	20,5	99	17,0
29	97,0	66	86,0	100	83,0
30	55,5	67	92,0	101	32,0
32	75,5	68	91,0	102	31,5
33	30,5	69	23,5	103	18,5
34	18,0	70	60,5	104	92,0
36	79,0	71	90,5	106	71,0

After the data selection process, the next step is to visualize the distribution plot of the data that has been selected. The results of the data plot can be seen in Figure 2. In Figure 2 we cannot easily see the outliers. However, as an evaluation, the value of outliers on learning outcomes is very important to obtain.

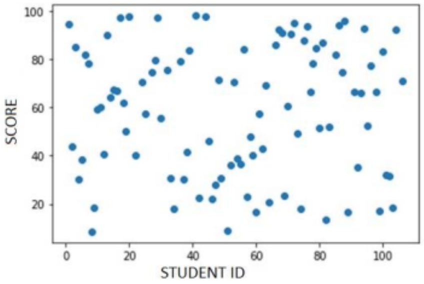

Figure 2. The Plot of Data Selection Results

Here are examples of Outlier Detection Steps using K-Means with Minkowski-Chebyshev Distance

Step 1: Calculate the cluster center of the initial matrix that is already raised at the start of the centroid is 9.0

Step 2: Defining the Minkowski-Chebyshev distance by using equation (3)

Step 3: Determining the distance between the early centroid

[6] J. B. Macqueen, "Some Methods For Classification And Analysis Of Multivariate Observations," Symp. Math. Stat. Probab., vol. 1, pp. 281–297, 1967, doi: 10.1007/s11665-016-2173-6.

[7] V. Bhatt, M. Dhakar, and B. K. Chaurasia, "Filtered Clustering Based on Local Outlier Factor in Data Mining," Int. J. Database Theory Appl., vol. 9, no. 5, pp. 275–282, 2016, doi: 10.14257/ijdta.2016.9.5.28.

[8] T. Christopher, "A Study of Clustering Based Algorithm for Outlier Detection in Data streams," no. March, pp. 194–197, 2015.

[9] Y. Erdem and C. Ozcan, "K-Means Clustering on Apache Spark," no. 7, pp. 86–90, 2017.

[10] C. Sumithiradevi and M. Punithavalli, "Enhanced K-Means with Greedy Algorithm for Outlier Detection," Int. J. Adv. Res. Comput. Sci., vol. 3, no. 3, pp. 294–297, 2012.

[11] J. Ren, "A detection algorithm of customer outlier data based on data mining technology," vol. 33, no. Febm, pp. 272–278, 2017, doi: 10.2991/febm-17.2017.35.

[12] J. Han, M. Kamber, and J. Pei, Data Mining Consepts And Techniques, 3rd ed. San Fransisco: Morgan Kaudmann, 2012.

[13] M. Kuppusamy and K. S. Kannan, "Comparison of methods for detecting outliers," Int. J. Sci. Eng. ..., no. January 2013, 2013, [Online]. Available: https://scholar.google.co.in/scholar?hl=en&q=manoj+and+senthamarai+kannan&btnG=#0.

[14] B. Iglewicz and D. C. Hoaglin, How to Detect and Handle Outliers, 16th ed. United States of America §: ASQC Quality Press Publications Catalog, 1993.

[15] B. Joyce and M. Weil, "Attaining concepts: The basic thinking skills," Model. Teach., pp. 161–178, 2003.

[16] W. Sanjaya, Strategi Pembelajaran Berorientasi Standar Proses Pendidikan. Jakarta: Kencana Perdana Media Group, 2008.

[17] M. Vahdat, L. Oneto, D. Anguita, M. Funk, and M. Rauterberg, "A Learning Analytics Approach to Correlate the Academic Achievements of Students with Interaction Data from an Educational Simulator," Springer Int. Publ. Switz., pp. 352–366, 2015, doi: 10.1007/978-3-319-24258-3.

HASIL CEK_ICAIBDA 21 Paper

ORIGINALITY REPORT

8%	5%	6%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	link.springer.com Internet Source	1%
2	Submitted to Trinity College Dublin Student Paper	1%
3	journal.uad.ac.id Internet Source	1%
4	Song, H.. "Attack-resilient time synchronization for wireless sensor networks", Ad Hoc Networks, 200701 Publication	1%
5	Sugiyarto Sugiyarto, Joko Eliyanto, Nursyiva Irsalinda, Zhurwahayati Putri, Meita Fitrianawat. "A Fuzzy Logic in Election Sentiment Analysis: Comparison Between Fuzzy Naïve Bayes and Fuzzy Sentiment using CNN", JTAM (Jurnal Teori dan Aplikasi Matematika), 2021 Publication	1%
6	Zhao-yu Shou, Meng-ya Li, Si-min Li. "Outlier detection based on multi-dimensional	1%

	clustering and local density", Journal of Central South University, 2017 Publication	
7	Sugiyarto Surono, Rizki Desia Arindra Putri. "Optimization of Fuzzy C-Means Clustering Algorithm with Combination of Minkowski and Chebyshev Distance Using Principal Component Analysis", International Journal of Fuzzy Systems, 2020 Publication	<1 %
8	www.mdpi.com Internet Source	<1 %
9	essayy.com Internet Source	<1 %
10	www.researchgate.net Internet Source	<1 %
11	journal.utem.edu.my Internet Source	<1 %
12	Junli Li, Jifu Zhang, Xiao Qin, Yaling Xun. "Feature grouping-based parallel outlier mining of categorical data using spark", Information Sciences, 2019 Publication	<1 %
13	Mushtaq Hussain, Wenhao Zhu, Wu Zhang, Syed Muhammad Raza Abidi, Sadaqat Ali. "Using machine learning to predict student	<1 %

	difficulties from learning session data", Artificial Intelligence Review, 2018 Publication	
14	Retno Dwi Suyanti, Deby Monika Purba. "The implementation of discovery learning model based on lesson study to increase student’s achievement in colloid", AIP Publishing, 2017 Publication	<1 %
15	Xiaochun Wang, Xiali Wang, Mitch Wilkes. "New Developments in Unsupervised Outlier Detection", Springer Science and Business Media LLC, 2021 Publication	<1 %
16	Mehrnoosh Vahdat, Luca Oneto, Davide Anguita, Mathias Funk, Matthias Rauterberg. "Chapter 26 A Learning Analytics Approach to Correlate the Academic Achievements of Students with Interaction Data from an Educational Simulator", Springer Science and Business Media LLC, 2015 Publication	<1 %
17	TINGSONG DU, HAO WANG, MUHAMMAD ADIL KHAN, YAO ZHANG. "CERTAIN INTEGRAL INEQUALITIES CONSIDERING GENERALIZED m-CONVEXITY ON FRACTAL SETS AND THEIR APPLICATIONS", Fractals, 2019 Publication	<1 %

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		