# Analysis of Representative Values in Clustering using the CURE Algorithm

**Umaya Ramadhani Putri Nst, Sutarman, Pahala Sirait**

***Abstract:*** ***The collection of data that everyone has on earth has a fully agreed upon value of knowledge. Analysis of a collection of data that can accommodate a long processing time, for this we need an algorithm that can provide a comparison of the acceleration of the analysis process. One process of data analysis is clustering, which is a process of grouping large amounts of data so that it is easy to understand. One of the algorithms in the clustering process is CURE (Clustering Using Representative) where CURE random sample-based data bases partition the data using representative points called representative points. Sample-based process will provide better processing time acceleration because it will only be done on the data collection, not the whole data. This representative point determines the processing time of the testing carried out in the input. Values, representative values, and shrinkage values will provide a faster settlement process for the values inputted according to the correct conditions.***

***Keywords*** **:** ***Representative, CURE algorithm***

## I. INTRODUCTION

The collection of data contained on this earth is a collection of all data possessed by everyone. Data from one person can have large amounts of data with large data variants too. The data set will eventually become a pile of data that can produce important information from each who has it.

A collection of data that accumulates can be extracted into a pattern structure or knowledge automatically when an analysis of the data is carried out. To that end, one of the data analysis is to cluster data so that data can be better understood by humans and have a value of knowledge.

Clustering is a grouping stage for analyzing large amounts of data [1]. This analysis is able to divide the clusters of data so that data is more easily understood. One algorithm for clustering is the CURE algorithm, Clustering Using Representative. CURE is an algorithm that uses a hierarchical method that combines random sampling-based data and then partitions the data so that the process is carried out only on the happiness of the data [1]. For this reason, CURE can be used on large amounts of data. CURE is an algorithm that determines a representative point as a reference point to form several clusters. This representative point is called the representative point which was initially determined by the input value of the representative point. The number of values from the representative point has an influence on the time the process is traveled, whether it will be able to speed up or slow- down in the analysis process. For this reason, an analysis of the value of the representative point input at the beginning and how to overcome it if it gives a long processing time.

## II. RELATED RESEARCH

This description begins with research by Manjula and Nandakumar (2018) which uses the CURE algorithm to identify educational data in data mining regarding engineering performance of weak engineering students. The educational data to be analyzed has many attributes so that this study performs irrelevant attribute reduction using the method of Reduction by Dimensionality Reduction. The resulting time runs for 11,929215 s.

Previously, in 2005 a study was carried out by Yin (2005) by classifying in large datasets using the BIRCH algorithm where CF-Tree was used to determine sub-clusters. The cluster formed is stored like a leaf node, then uses k-prototype to correct the node if the cluster is not balanced.

The next four years of research conducted by Meng, Song and Wang (2009) who proposed a new algorithm to handle the clustering process in complex and massive data effectively and efficiently. The proposed algorithm is an algorithm that is adopted from the grid shape and then grouped into clusters by looking at the density of the grid shape.

Eick, Zeidat and Vilalta (2004) carried out the research by editing in advance the dataset used to improve the classification accuracy. One example is removing an object from the training set and being given a decision limit.

Rani, Manju and Rohil (2014) conducted a study comparing the results of clustering using the CURE and BIRCH algorithm with WEKA data 3.6.9. the results obtained are the CURE algorithm provides the best cluster results than the BIRCH algorithm, but in terms of time, BIRCH gives a faster time than the CURE algorithm.

**Revised Manuscript Received on June 29, 2020**.
***** Correspondence Author
**Umaya Ramadhani Putri Nst**, Bachelor of Information Technology, Universitas Sumatera Utara, Indonesia. Email: umiequ@gmail.com.
**Sutarman**, Magister of Applied Probability and Statistics, Northern Illionis University, Amerika. Email: sutarman@usu.co.id.
**Pahala Sirait**, Magister of Computer Science, Universitas Indonesia, Indonesia. Email: sirpahala@yahoo.com.

## III. PROPOSED METHOD

### A. Min-Max Normalization Method

The Min-Max method is a normalization method by performing a linear transformation of the original data [2]. The formula is as follows:

$$new\ data = \frac{(data-min)*(newmax-newmin)}{(max-min)} + newmin \quad (1)$$

In this case:

$new\ data$ = Normalized data
$min$ = Minimum value of data per column
$max$ = Maximum value of data per column
$newmin$ = The minimum we set
$newmax$ = The maximum we set

### B. CURE Algorithm

CURE (Clustering Using Representative) Algorithm is an algorithm that uses a hierarchical method that combines random sampling-based data and then partitiones the data so that the process is carried out only on the happiness of the data [1]. This algorithm is created to identify data that will form random clusters with wide variations [3].

CURE presents each cluster from certain points that are scattered by shrinking the cluster center using linear space so as to produce a faster process [4]. The main purpose of CURE is to look for these representative points called representative points in order to get good cluster results [5]. Although it produces a fast time process in the clustering process, CURE still includes algorithm which has a bad time complexity with values $O(n^2 logn)$.

The stages of CURE can be seen in Figure 1 below, namely the flowchart of CURE algorithm.

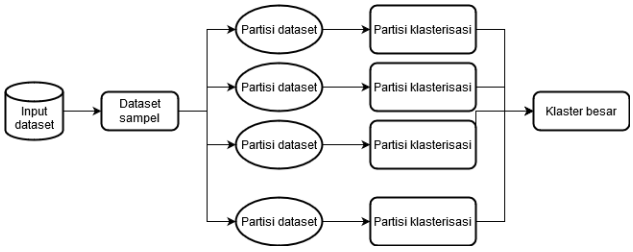


**Fig 1. Flowchart CURE [1]**

Following is the CURE algorithm process [6]:

1) Take a random sample of data $n$ from the dataset.
2) Partitioning $p$ to the sample $n$ becomes a size $\frac{n}{p}$, where the value $p$ = 2, here will form two initial partitions by having $\frac{n}{p}$ the data contents of each cluster.
3) Then each initial partition is partitioned back into a size $\frac{n}{pq}$, where $q$ > 1.
4) Establish the number of points as a representative point for each cluster.
5) The representative points then shrink with the given shrink value forming a new representative point.
6) The two clusters with the closest distance value are then combined.
7) After that, a representative point is chosen as the representative to represent the search for new clusters.
8) Cluster merge stops after $k$ cluster targets meet.

## IV. RESULT AND DISCUSSION

### A. Research Data

The data in this study used a dataset taken from the UCI Machine Learning dataset, which is data from a customer's credit card in Taiwan from April to September 2005. The dataset consists of 30,000 data and 24 attributes. The dataset will be normalized using the min-max normalization method into a balanced range of values to facilitate the calculation of trials.

Following are examples of data that have not been normalized and after normalization, 30 data are taken from 30,000 data.

**Tabel- I: Dataset Before Normalized**

| *No* | $X_1$ | $X_2$ | $X_3$ | $X_4$ | ... | $X_{23}$ |
|---|---|---|---|---|---|---|
| *1* | *50000* | *1* | *1* | *2* | ... | *0* |
| *2* | *230000* | *2* | *1* | *2* | ... | *0* |
| *3* | *50000* | *1* | *2* | *2* | ... | *716* |
| *4* | *100000* | *1* | *1* | *2* | ... | *2504* |
| *5* | *500000* | *2* | *2* | *1* | ... | *51582* |
| *6* | *500000* | *1* | *1* | *1* | ... | *768* |
| ... | ... | ... | ... | ... | ... | ... |
| *30* | *200000* | *2* | *1* | *2* | ... | *0* |

**Tabel- II: Dataset After Normalized**

| *No* | $X_1$ | $X_2$ | $X_3$ | $X_4$ | ... | $X_{23}$ |
|---|---|---|---|---|---|---|
| *1* | *0,08* | *0,00* | *0,00* | *0,50* | ... | *0,00* |
| *2* | *0,45* | *1,00* | *0,00* | *0,50* | ... | *0,00* |
| *3* | *0,08* | *0,00* | *0,25* | *0,50* | ... | *0,01* |
| *4* | *0,18* | *0,00* | *0,00* | *0,50* | ... | *0,05* |
| *5* | *1,00* | *1,00* | *0,25* | *0,00* | ... | *1,00* |
| *6* | *1,00* | *0,00* | *0,00* | *0,00* | ... | *0,01* |
| ... | ... | ... | ... | ... | ... | |
| *30* | *0,39* | *1,00* | *0,00* | *0,50* | ... | *0,00* |

### B. Result and Discussion

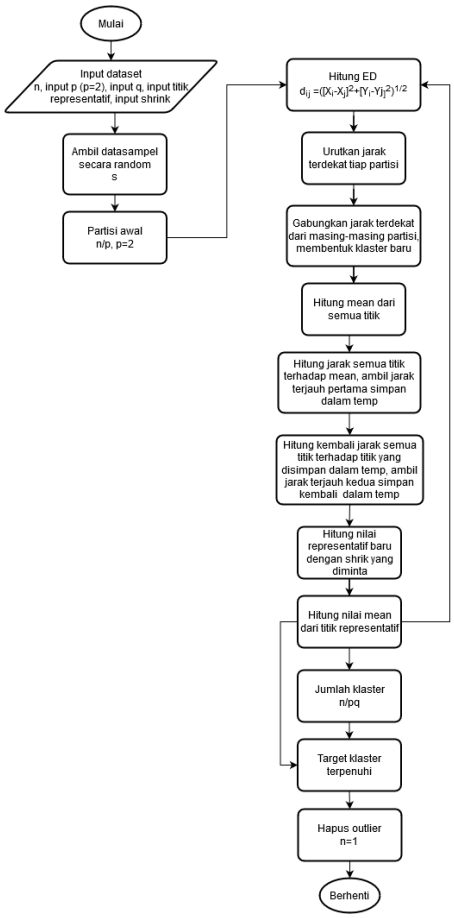


**Fig. 1. Flowchart design CURE**

In Figure 1 above, a flowchart picture of the CURE algorithm is performed in this test. Starting from the input stage some values then produce output as a result of the tests conducted.

Testing the CURE algorithm in this study is to test the requested input values. In the CURE process, there are some inputs that we must know, namely input values $q$ and values that are representative points. We can analyze the input values so we know the strengths and weaknesses of each input value in the CURE process. Input value$q$ is the value used for the second partition stage of the clustering process so that it can further reduce the calculation process but slow down the process. Representative value is a value that is used as a reference point that will be used for the boundaries of a cluster of clusters, the representative value given is also able to minimize the calculation process but can slow down the process. The following table 3 shows the results of the test by testing the input value $q$.

**Tabel- III: Input Value Testing Results $q$**

| *Input Value* $q$ | *Processing Time* | *Number of Representative Points* | *Number of Clusters Produced* |
|---|---|---|---|
| *2* | *0,393 s* | *2* | *3* |
| *3* | *0,172 s* | *2* | *3* |
| *4* | *0,155 s* | *2* | *3* |
| *5* | *0,094 s* | *2* | *3* |
| *6* | *0,068 s* | *2* | *3* |
| *7* | *0,067 s* | *2* | *3* |
| *8* | *0,067 s* | *2* | *3* |
| *9* | *0,068 s* | *2* | *3* |
| *10* | *0,065 s* | *2* | *3* |

From Table III above, it can be concluded that the greater the input value $q$ is given, the shorter the processing time because the more partitions are formed so that the process runs simultaneously with smaller calculations. From the test results it can be seen the number of clusters that remain stable without changing.

**Tabel- IV: Test Results Representative Input Value**

| *Number of Representative Points* | *Processing Time* | *Number of Clusters Produced* |
|---|---|---|
| *2* | *0,2 s* | *3* |
| *3* | *0,125 s* | *3* |
| *4* | *0,147 s* | *3* |
| *5* | *0,179 s* | *3* |
| *6* | *0,240 s* | *3* |
| *7* | *0,265 s* | *3* |
| *8* | *0,271 s* | *3* |
| *9* | *0,377 s* | *3* |
| *10* | *0,473 s* | *3* |

The test results in Table IV, show that when the number of representative points is getting bigger, the process time is getting longer, concluding that CURE has a weakness in giving the best number of representative values. Nevertheless, it still provides a stable number of clusters that does not change. Here it is seen that a trial is needed on the two inputs in order to find the right input values to provide better processing results. The test results in Table V below will minimize the above weaknesses, namely by providing appropriate input values and representative values by trying several tests. After a number of tests have been conducted, it can be concluded that if the values$q$ > representative values inputted, will result in shorter processing time.

**Tabel- V: Modification $q$ and Representative Testing Results**

| *Input Value $q$* | *Number of Representative Points* | *Processing Time* | *Number of Clusters Produced* |
|---|---|---|---|
| *5* | *2* | *0,17 s* | *3* |
| *5* | *3* | *0,155 s* | *3* |
| *5* | *4* | *0,121 s* | *3* |

In the process of clustering, the CURE algorithm is known as a shrink value. This shrink value is between 0 and 1. The shrink value here is a value to narrow the representative point that is used as a reference point to other points.

For changes that occur when the shrink value is inputted, if the results of the above conditions, namely, the value> representative value, plus the shrink value is enlarged to reach a value of 1, then, the processing time obtained will be even faster than before. we can see in Table VI, Table VII and Table VIII below.

**Tabel- VI: Testing Results Input Value $q$=5 R=2**

| *Value of Shrink* | *Processing Time* | *Number of Clusters Produced* |
|---|---|---|
| *0,3* | *0,187 s* | *3* |
| *0,5* | *0,08 s* | *3* |
| *0,8* | *0,075 s* | *3* |
| *1* | *0,072 s* | *3* |

**Tabel- VII: Testing Results Input Value $q$=5 R=3**

| *Value of Shrink* | *Processing Time* | *Number of Clusters Produced* |
|---|---|---|
| *0,3* | *0,115 s* | *3* |
| *0,5* | *0,094 s* | *3* |
| *0,8* | *0,157 s* | *3* |
| *1* | *0,098 s* | *3* |

**Tabel- VIII: Testing Results Input Value $q$=5 R=4**

| *Value of Shrink* | *Processing Time* | *Number of Clusters Produced* |
|---|---|---|
| *0,3* | *0,5 s* | *3* |
| *0,5* | *0,159 s* | *3* |
| *0,8* | *0,128 s* | *3* |
| *1* | *0,13 s* | *3* |

## V. CONCLUSION

In this study, it can be concluded that the CURE algorithm can cluster the data in large amounts with large variants with stable cluster results reaching k targets. This study also shows that in the trial analysis of changes in values$q$, representative values and shrink values provide a faster time change where the conditions are as follows:

1) When the value$q$ is entered, the processing time becomes shorter.
2) When the value of a representative value is entered, the processing time becomes longer
3) When the value$q$> representative value, the processing time is faster.
4) If the 3rd point plus the shrink value is enlarged to reach 1, the processing time will be even faster than before.

## REFERENCES


1. Shirkhorsidi, A.S., Aghabozorghi, S., Wah, T.Y., & Herawan, T. 2014. Big data Clustering. International Conference Computational Science and Applications, pp. 707-720.
2. Rani, Y., Manju & Rohil, H. 2014. Comparative Analysis of BIRCH and CURE Hierarchical Clustering Algorithm using WEKA 3.6.9. Computer Science Engineering and Application 2: 25-29.
3. Jian, S., Pang, G., & Cao, L. 2018. CURE: Flexible Categorical Data Representation by Hierarchical Coupling Learning. IEEE Transactions on Knowledge and Data Engineer. Vol. 14. No.8
4. Eick, Christoph F., Zeidat, N., & Vilalta, R. 2004. Using Representative-Based Clustering for Nearest Neighbor Dataset Editing. Proceedings of the fourth IEEE International Conference on Data Mining, pp. 2142-2145.
5. Putra, A.K.P., Purwanto, Y., & Novianty, A. 2015. Analisis Sistem Deteksi Anomali Traffic Menggunakan Algoritma CURE dengan Koefisien Silhoutte dalam Validasi Clustering. Procedings of engineering, pp. 3837-3842.
6. Han, J., Kamber, M., & Pei, J. 2012. Data Mining: Concepts and Techniques Third. Elsevier: USA.
7. Manjula, V. & Nandakumar, A.N. 2018. An Effective Cure Clustering Algorithm in Education Data Mining Techniques to Valuate Student's Performance. International Journal of Applied Engineering Research 10: 7493-7498.
8. Meng, H.-D., Song, Y.-C., Song, F.-Y., & Wang, S. L. 2009. Clustering for Complex and Massive Data. International Conference on Information Engineering and Computer Science, pp. 4244-4994.
9. Rani, Y., Manju & Rohil, H. 2014. Comparative Analysis of BIRCH and CURE Hierarchical Clustering Algorithm using WEKA 3.6.9. Computer Science Engineering and Application 2: 25-29.
10. Yin, J., Tan, Z., Ren, J., & Chen, Y. 2005. An Efficient Clustering Algorithm Mixed Type Attributes in Large Dataset. Proceedings of the Fourth International Conference on Machine Learning and Cybernatics, pp. 7803-9091.


## AUTHORS PROFILE

**Umaya Ramadhani Putri Nst,** Graduate school of Information Technology, Universitas Sumatera Utara, Indonesia. The first research title is "The Design of Vehicle Laying Model in Ticket Booking Application Ship" on International Journal of e-Ducation, e-Business, e-Management, and e-Learning.

**Sutarman**, Magister of Applied Probability and Statistics, Northern Illionis University, Amerika, 1994.

**Pahala Sirait**, Magister of Computer Science, Universitas Indonesia, Indonesia, 2004.