Bigdata Clustering using X-means method with Euclidean Distance

N Zendrato1*, H W Dhany2, N A Siagian1* and F Izhari2*
1 Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia.
2 Universitas Pembangunan Pancabudi Medan, Indonesia.

*Email: niskarto@usu.ac.id, hdnany@dosen.pancabudi.ac.id, novriadi95@yahoo.com, fahmi_iizhari@dosen.pancabudi.ac.id

Abstract. Centroid is the central point of data in the grouping process, it is necessary to analyze the centroid in determining the initial value in the initial clustering process. So it is used as a cluster center point in the X-Means algorithm clustering process. Determine cluster center points or centroid, measure the performance of the X-Means algorithm with range cluster parameters by measuring distances between centroid for a fast and efficient way to group unstructured data, and to speed up the model construction process and divide several centroid in half to match the data as a test tool for the analysis of the X-Means method. From testing using the X-Means algorithm with the determination of the number of Centroid clusters carried out by modifying the X-Means method to do some determination of the centroid to get the results of 11 iterations. From the results of these tests produce good cluster members the level of similarity of data with other data and in determining the number of clusters, using the modification of the Euclidean distance method, get better results of the similarity level of each member compared to randomly determining the number of clusters with several iterations.

Keywords: Clustering, Centroid, X-Means, Iteration, and Accuracy.

1. Introduction

Centroid clusters are randomly selected through a number of K-clusters. The algorithm divides the data provided into K-clusters, each of which has its own cluster membership and assigns each data point to the nearest center of mass. Then recompile the centroid using the current cluster association and if the grouping does not merge, the process will be repeated several times. X-means clustering is a variation of K-means clustering which treats cluster allocations by trying repetitive partitions and maintaining optimal outcome separation, until several criteria are reached. The goal of the X-mean cluster is to do intrinsic grouping in an unlabeled data set. Provides a fast and efficient way to classify unstructured data, the use of concurrency by accelerating the process of model construction and usage.

Some researchers have looked for the problem of k-means clustering and there are many approaches to accelerate k-means. But several methods have been introduced to scalability and reduce the time complexity of the k-means algorithm. [1] The purpose of this method is to divide several centroids in two to match the data achieved. The X-means algorithm has proven to be more efficient.
than k-means in grouping data more accurately. This method does not have any disadvantage, in separating the selection of many centroids when the data is not completely spherical.

The process of grouping data into clusters or grouping so that data in one cluster has a maximum level of similarity and between clusters has a minimum similarity is called Clustering. Clustering is divided into 2 approaches in its development namely clustering partitioning and hierarchical approaches [3].

Grouping x-means that are strong in combining the formation of classes from classifications, x-means has the optimization of clustering models with little time. X-means is can give very satisfying results in one class problem. However, checking different criteria in measuring the similarity of models for each cluster should be more focused [2].

Big data is data that has a large size and its volume will continue to grow, which there are various types of data, are formed continuously at a certain speed, and must be processed at a certain speed as well. Big data includes data that has been very difficult to be collected, stored and managed or analyzed using the usual database system because its volume continues to multiply. In terms of technology, there will emerge the importance of the ability to process big data [4].

Large amounts of data are better used as a whole because of the possibility of correlations at larger amounts, correlations that can never be found in the data are analyzed in separate sets or smaller sets. Larger amounts of data provide better output but also work with limited processes [5, 6].

2. Research methods

X-means clustering is used to solve one of the main weaknesses of K-means clustering, namely the need for prior knowledge about the number of clusters (K). In this method, the true value of K is estimated in an unsupervised way and only based on the data set itself. \(K_{\text{max}}\) and \(K_{\text{min}}\) as upper and lower limits for the possible values of X. In the first step of X-means grouping, knowing that now \(X = X_{\text{min}}\), X-means finding the initial structure and centroid. In the next step, each cluster in the structure estimated to be treated as a parent cluster, which can be divided into two groups. The score helped decide is a better representation for sample data. Clusters provide more accurate distribution of samples. As a result, either the parent parent will be replaced by centroid, or the algorithm will remain centroid and leave it. Then, the structure will be built or updated based on choices. This procedure will be continued for all clusters in the initial structure until now the estimated number of clusters is greater than the \(X_{\text{max}}\) or algorithm converging to the best structure. This algorithm can be too slow because it needs to restart X-means for each cluster split. To solve this problem, apply the kd-tree from the data set which naturally reduces the number of nearest neighbor requests for K-means (Pelleg, 2000).

Testing patterns using data as training data, namely the formation of test data clusters, the first time is euclidean distance, then proceed to the chebychev distance method, corellation similarity, cosine similarity, dice similarity, dynamic time, warping distance with testing of centroid 2,3,4, and 5.

3. Result and discussion

Tests conducted by researchers that use data sets Majors High School Students which include: No, Class, ID, Mathematics, Biology, Physics, Economics, Geography, Chemistry, Sociology, IQ. In the X-Means method with the determination of the centroid value, as well as an analysis of some of the methods applied also can get results in research by clustering data into several clusters in the data set of Majors for Senior High School Students as an initial step of analysis by conducting the best data grouping on the X method -Means by conducting several tests of determining the centroid value up to 8 times testing.

3.1. Model used

In the predetermined dataset the model of the method to be used is X-Means with a chart of the RapidMiner software used, which can be seen as follows:
3.2. Data testing
In testing the data, clustering analysis using software, RapidMiner, is performed. The dataset was obtained from the results of research conducted at school by obtaining student grades. The data being tested was aimed at studying the various clusters contained in the program as well as the differences resulting from each method tested. Clustering Process with X-Means
This test aims to see the effect of the number of documents, number of clusters, and clustering methods in classifying documents. Then the clustering process is done using X-Means, while the stages are as follows:

1. Initialization of cluster number (C = 2) with cluster center 1 and cluster center 2 is taken randomly, while the initial cluster center can be seen in table 1 below:

| ID Students | X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 | Target |
|-------------|----|----|----|----|----|----|----|----|--------|
| 85          | 75 | 77 | 72 | 73 | 70 | 85 | 91 |   | IPA    |
| 165         | 70 | 80 | 70 | 69 | 78 | 78 | 90 |   | IPS    |

2. Calculate the distance of each data to each of the initial cluster centers in table 1 with the following equation:

\[ d(x_2, x_1) = \|x_2 - x_1\|^2 = \sum_{j=1}^{n} (x_{2j} - x_{1j})^2 \]

Information:
- \( x_{1j} \): i data from j-attribute
- \( x_{2j} \): i cluster center of the j attribute
- \( n \): number of data characteristics (attributes)
- \( d(x_2, x_1) \): Euclidean Distance, the distance between data at points \( x_1 \) and \( x_2 \)

The following is the calculation of the distance between the data centers of the cluster in table 1:

Iteration 1:
1st Student ID:

\[ C_1 = \sqrt{(89 - 75)^2 + (96 - 77)^2 + (96 - 72)^2 + (93 - 73)^2 + ... + (127 - 91)^2} = 58.06 \]

\[ C_2 = \sqrt{(89 - 70)^2 + (96 - 80)^2 + (96 - 70)^2 + (93 - 69)^2 + ... + (127 - 90)^2} = 62.63 \]
ID_Student-2:

\[ C_1 = \sqrt{(88 - 75)^2 + (95 - 77)^2 + (95 - 72)^2 + (90 - 73)^2 \ldots + (115 - 91)^2} = 49.50 \]

\[ C_2 = \sqrt{(88 - 70)^2 + (95 - 80)^2 + (95 - 70)^2 + (90 - 69)^2 \ldots + (115 - 90)^2} = 53.71 \]

ID_Student-3:

\[ C_1 = \sqrt{(83 - 75)^2 + (93 - 77)^2 + (94 - 72)^2 + (87 - 73)^2 \ldots + (109 - 91)^2} = 41.82 \]

\[ C_2 = \sqrt{(83 - 70)^2 + (93 - 80)^2 + (94 - 70)^2 + (87 - 69)^2 \ldots + (109 - 90)^2} = 46.04 \]

Do the same thing until all student data is obtained each distance from the center of cluster 1 (C1) and center of cluster 1 (C2). The results of the calculation of the distance between data can be seen in Table 2.

| ID Student | Distance | Distance | Cluster |
|------------|----------|----------|---------|
|            | C1       | C2       | Nearest |         |
| 1          | 58.06    | 62.63    | 58.06   | 1       |
| 2          | 49.5     | 53.71    | 49.5    | 1       |
| 3          | 41.82    | 46.04    | 41.82   | 1       |
| 4          | 45.56    | 48.98    | 45.56   | 1       |
| 5          | 40.04    | 43.93    | 40.04   | 1       |
| 6          | 39.34    | 43.05    | 39.34   | 1       |
| 7          | 33.53    | 38.43    | 33.53   | 1       |
| 8          | 31.5     | 34.66    | 31.5    | 1       |
| 9          | 38.69    | 41.74    | 38.69   | 1       |
| 10         | 35.62    | 40.72    | 35.62   | 1       |
| ...        | ...      | ...      | ...     | ...     |
| 167        | 12.65    | 9.95     | 9.95    | 2       |

To see the full breakdown of Data Distance to Cluster Center (Iteration 1), see Appendix 2.

3. Determine the new cluster center by calculating the average attribute value of each data included in the cluster in table 2. The new cluster values can be seen in table 3 below:

| Cluster | X1   | X2   | X3   | X4   | X5   | X6   | X7   | X8   |
|---------|------|------|------|------|------|------|------|------|
| C1      | 79.65| 79.38| 79.14| 79.97| 77.53| 77.84| 82.96| 100.97|
| C2      | 71.02| 71.6 | 68.81| 72.35| 69.73| 75.23| 74.48| 91.88 |

4. Back to step 2 until there is no data to move the cluster or to the maximum iteration limit. In this study, the convergence of student majors data was obtained during the 11th iteration. The final cluster center values found can be seen in table 4 below:
Table 4. Initialization of the Final Cluster Center (11th Iteration).

| Cluster | X1  | X2  | X3  | X4  | X5  | X6  | X7  | X8  |
|---------|-----|-----|-----|-----|-----|-----|-----|-----|
| C₁      | 83.61 | 85.56 | 84.98 | 83.21 | 80.69 | 80.77 | 85.37 | 105.53 |
| C₂      | 73.04 | 71.88 | 70.57 | 74.29 | 71.8  | 74.82 | 77.33 | 93.77  |

If it is tested using the Rapidminer application, the results of manual calculations with Rapidminer are the same as the following display:

![Figure 2. Result Cluster.](image)

Table 5. Cluster Results using X-Means.

| ID Student | Cluster | Majors |
|------------|--------|--------|
| 1          | 1      | IPA    |
| 2          | 1      | IPA    |
| 3          | 1      | IPA    |
| 4          | 1      | IPA    |
| 5          | 1      | IPA    |
| 6          | 1      | IPA    |
| 7          | 1      | IPA    |
| 8          | 1      | IPA    |
| 9          | 1      | IPA    |
| 167        | 2      | IPS    |

4. Conclusion
The testing using the X-Means algorithm with the determination of the number of centroid clusters, the X-Means method is modified by using Euclidean distance to get the results of 5 clusters. From the results of these tests produce a good cluster member level similarity of data with other data. In determining the number of clusters, using the modification of the Euclidean distance method gets
better results for the level of similarity of each of its members compared to determining the number of clusters randomly.

References

[1] Pelleg D and Moore A. 2000. X-means: Extending K-means with Efficient Estimation of The Number of Clusters. Proceedings of the Seventeenth International Conference on Machine Learning, Palo Alto, CA.

[2] Krawczyk Bartosz and Michał Woźniak. 2013. Pruning Ensembles of One Class Classifiers with X-means Clustering. Wrocław University of Technology.

[3] Poteras, C.M., Mihăescu, M.C. & Mocanu, M. 2014. An optimized version of the kmeans clustering algorithm. Proceedings of the 2014 Federated Conference on Computer Science and Information Systems, pp. 695–699.

[4] Sujana Aprianti Putri. 2013. Utilizing Big Data to Detect Emotions. Computer Engineering School of Electrical and Information Engineering Bandung Institute of Technology.

[5] Ularu Elena Geanina, Puican Florina Camelia, Apostu Anca, Velicanu Manole. 2012. Perspectives on Big Data and Big Data Analytics. Database Systems Journal vol. III, no. 4.

[6] Fariza, A., Arifin, A.Z., Astuti, E.R. and Kurita, T., Segmenting Tooth Components in Dental X-Ray Images Using Gaussian Kernel-Based Conditional Spatial Fuzzy C-Means Clustering Algorithm.