Analysis of Centroid Cluster in X-Means Clustering in Data Classification: Power Absorb Oxygen

Sardo Pardingotan Sipayung[1], Poltak Sihombing[1] and Sutarman[2]
[1]*Department of Computer Science, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia*
[2]*Department of Information Technology, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara*

Keywords: Oxygen, Cluster, Centroid, X-Mean.

Abstract: On gardens city of Medan, there are type different trees. On every tree have power absorbency oxygen and work issue oxygen every day. It will be grouping the tree data that issued oxygen with use X-Means method on Clustering algorithm. Then in research, an analysis to centroid that is point data center inside process grouping, then need to an analysis centroid in determining gift value early to process the beginning of clustering. So that data was used as point center cluster on process X-Means clustering algorithm.

1 INTRODUCTION

Centroid *cluster* selected in a manner random through a number of *K-cluster*. Algorithm share the data provided to in *K-cluster*, respectively have membership *cluster* own and set every data point to center mass closest. Then compile reset it centroid use association cluster when this and if grouping not fused, the process will be repeat to several times. *X-means clustering* is variation from *K-means clustering* treat allocation *cluster* with try partition over and over and keep separation optimal results, arrive some criteria achieved. *X-mean cluster* with do grouping intrinsic in a data set that is not labeled. Giving fast way and efficient for grouping data that doesn't structure, usage *concurrency* with speed up process model and construction use.

Point center *cluster* or *centroid* is a point early start grouping in the *cluster* on algorithm *K-Means.* Data grouping is done with calculating distance closest with point center initial *cluster* as point central information every group or *cluster*. However on its application, determination point center initial *cluster* this is what become weakness from algorithm *K-Means*. This caused because not there is an approach used to choose and determine point center *cluster*. Point center *cluster* selected in a manner just any or random from a set of data. The results *clustering* from algorithm *K-Means* often less optimal and not maximum in every experiment conducted. By because that, can say it that well bad the results *clustering*, very depend on point center *cluster* or *centroid* beginning (Baswade, 2013).

Some researchers have looked for the problem of k-means clustering and some have taken many approaches to accelerate k-means. But several methods have been introduced to scalability and reduce the time complexity of the k-means algorithm. (Pelleg, 2000) has proposed a method called X-means. The purpose of this method is to divide several centroids into two to match the data reached. The X-means algorithm has proven to be more efficient than k-means. This method does not have any disadvantages, based on the BIC (Bayesian Information Criterion) on the separation of many centroid selections when the data is not completely spherical.

2 RESEARCH METHODS

2.1 Clustering

Clustering is method classify or partition data inside a dataset. On basically *clustering* are something method for looking for and group data that has similarity characteristic *(similarity)* between one data with other data (Bhusare, 2014). The *Cluster* is a group data objects that have similarity one each other

Sipayung, S., Sihombing, P. and Sutarman, .
Analysis of Centroid Cluster in X-Means Clustering in Data Classification: Power Absorb Oxygen.
DOI: 10.5220/0008547601350137
In *Proceedings of the International Conference on Natural Resources and Technology (ICONART 2019)*, pages 135-137
ISBN: 978-989-758-404-6

inside of *cluster* and who doesn't have similarity to objects that are different *cluster*. Object will grouped to in one or more *cluster* so objects that are located in one *cluster* will have a high similarity between one with others. The objects will be grouped based on principle maximizing similarity object on *cluster* and maximizing inequality on a different *clusters.* Similarity object usually obtained from values attribute that explains data object, whereas data objects usually represented as a point in room multidimensional. Characteristics from every *cluster* not determined before, however pictured from data similarity grouped in inside it.

2.2 X-Means Clustering

X-means clustering is used for completely wrong the other weakness main from K-means clustering, that is the need knowledge previous about a number of clusters (K). In method this, value in fact from K estimated in something that isn't watched over way and only based on that data set alone.

Figure 1 Steps General In X-Means Grouping.

K_{max} and K_{min} as limit on and under for possible values from K. Step first X-Means grouping, knowing that when this is K = K_{min}, K-means find structure early and centroid. In step then, every cluster in the expected structure treated as parent cluster, which can divide to be two group children. Based on some criteria, which will explain in part next, we rate structure parents and children. Score help decide is person old is representations well for sample data or children. Cluster gives more distribution accurate on sample. As a result, a good parent will be replaced by centroid children, or algorithm will permanent person old centroid and leave children. Then, the new structure will be built or updated based on selection person old or children. Procedure this will next for all clusters inside structure early to when this estimated number of clusters to be bigger from max. K algorithm convergent to structure the best. Algorithm this can too slow because need run reset it K_{means} for every separation cluster. For resolve problem, apply kd-tree of data set that is natural reduce total demand neighbor closest for K-means (Pelleg, 2000).

3 IDENTIFICATION PROBLEMS

The centroid is point data center inside process grouping, then need to an analysis centroid in determine gift value early in process the beginning of clustering. So that used as point center *cluster* on process X-Means *clustering* algorithm.

4 RESULT AND DISCUSSION

The purpose of this study is to determine the center point of the cluster or centroid, measure the performance of the X-Means algorithm with range cluster parameters and compare the results of the X-Means algorithm accuracy with the k-means algorithm and by measuring the distance between centroids for fast and efficient ways to group unstructured data, and to speed up the process of construction of the model and divide some centroids into two to match the data achieved.

Results a reason about algorithm Clustering on X-Means method uses Power dataset Absorb Oxygen on Tree could be seen as the following of Cluster used:

Cluster 0: 104 items
Cluster 1: 3 items
Total number of items: 107

On the results analysis centroid, can be seen on a table the following:

Table 1: Analysis Centroid.

Attribute	Cluster_0	Cluster_1
Name of Tree	0.0	0.0
kg / year	-0.16	5.7
ton / year	-0.17	5.7

Table 2: Analysis Performance Vector.

entroid distance	-0.104
entroid distance cluster0	-0,013
entroid distance cluster1	-3,237
)avies Bouldin	-0.206

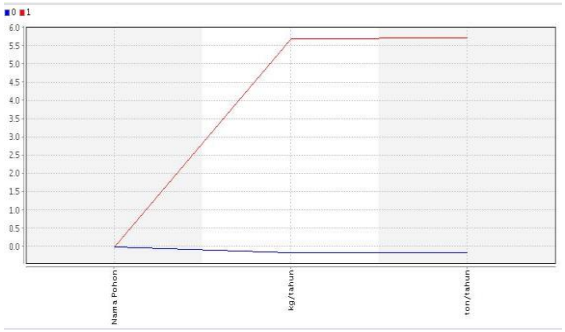

Figure 2: Cluster Graph.

5 CONCLUSIONS

Based on trial and analysis that has been done, obtained the results conclusion researcher as the following:

1. From measurements X-means accuracy has obtained the results namely: structure *clustering* obtained is nature *medium*.
2. Results measurement performance from the cluster, there are different distances between clusters 1 and 0.
3. More and more the size of the dataset used, then more and more the greater the value obtained but not change a number of clusters produced.
4. X-Means proved to have level Good accuracy compared with K-means with classifying type tree that has power absorbency oxygen.

REFERENCES

Baswade, A. M., Nalwade, P. S. 2013. Selection of Initial Centroids for K-Means Algorithm. International *Journal of Computer Science and Mobile Computing* (IJCSM) 2 (7): 161-164.

Bhusare, B. B., Bansode, S. M. 2014. Centroids initialization for K-means clustering using improved pillar algorithm. *Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, *3*(4), 1317-1322.

Pelleg, D., A. Moore. 2000. X-means: Extending K- effects with Efficient Estimation of The Number of Clusters. In *International Conference on Machine Learning*, Palo Alto, CA. 1, 727-734.

Krawczyk, B., Woźniak, M. 2015. Pruning Ensembles of One-Class Classifiers with X-means Clustering. In *Asian Conference on Intelligent Information and Database Systems*. 484-493.

Maimon, O., Last, M. 2001. *Knowledge Discovery and Data Mining*. Springer US. United States, 1st edition.

Poteras , C. M., Mihăescu, M. C., Mocanu, M. 2014. An optimized version of the kilometer clustering algorithm. In *Federated Conference on Computer Science and Information Systems*. 695–699.

Rose, JD 2016. An efficient association of rule based hierarchical algorithm for text clustering. *International Journal of Advanced Engineering Technology* 7 (4): 751 - 753.

Turban, E., E. Jay., Aronson., Liang Ting- Peng. 2005. *Decision Support System and Intelligent System*. Andi Offset