International Journal of Computer Science and Software Engineering (IJCSSE), Volume 4, Issue 6, June 2015
ISSN (Online): 2409-4285 www.IJCSSE.org Page: 160-164

Determining a Cluster Centroid of K-Means Clustering Using Genetic Algorithm

Hartono[1], Erianto Ongko[2] and Dahlan Abdullah[3]

[1, 2, 3] Department of Computer Science, University of Sumatera Utara, Medan, Indonesia

[1]*hartonoibbi@gmail.com,* [2]*erianto_ongko@yahoo.co.id,* [3]*dahlan@unimal.ac.id*

ABSTRACT

Clustering is a function of data mining that served to define clusters (groups) of the object in which objects are in one cluster have in common with other objects that are in the same cluster and the object is different from the other objects in different clusters. One method of clustering that can be used is the K-Means Clustering are included in the category of partitioning methods. One of the stages yan important in the K-Means Clustering is the cluster centroid determination, which will determine the placement of an object into a cluster based on the shortest distance between the object coordinate with cluster centroid. Genetic algorithms can be used in determining the initial value of the cluster centroid. the data set used in this study is the Iris data sets derived from the UCI Machine Learning Repository. Genetic algorithm is a heuristic search algorithm based on the idea of natural selection that Occurs in the process of evolution and genetic operations. This algorithm perform an intelligent search for a solution and have a broad spectrum of possible sollution. The determination of the initial value of the cluster centroid using genetic algorithms can provide better results than by using random numbers.

Keywords: *Clustering, K-Means Clustering, Cluster Centroid, Genetic Algorithm.*

1. INTRODUCTION

Clustering is a function of data mining that served to define clusters (groups) of the object in which objects are in one cluster have in common with other objects that are in the same cluster and the object is different from the other objects in different clusters [1]. The goal of clustering is to find a high-quality cluster where the distances between clusters is maximal and distance in the cluster is minimal [2]. One method of clustering that can be used is the K-Means Clustering are included in the category of partitioning methods [3]. In the K-Means clustering method will do the grouping objects into k groups or clusters. To do this clustering, k value must be determined in advance and the next step is to determine the cluster centroid [4]. One of the stages yan important in the K-Means Clustering is the cluster centroid determination, which will determine the placement of an object into a cluster based on the shortest distance between the object coordinate with cluster centroid [5]. Genetic Algorithm is a searching method used for choosing the best solution of the different problems, based on the mechanism of natural selection. That is, from the initial population, through several evolutionary steps, a set of new more appropriate solutions are achieved that led to the global optimal solution. This algorithm perform an intelligent search for a solution and have a broad spectrum of possible sollution. We can combine the max-min composition method to get an idea of the strength of the relationship between elements. Based on the strength of that relationship we can determine the interval membership function by using a genetic algorithm [6]. Genetic algorithms can be used in determining the initial value of the cluster centroid.

2. K-MEANS CLUSTERING

The K-Means Clustering algorithm is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. In 1967, Mac Queen [7] firstly proposed the K-Means algorithm. During every pass of the algorithm, each data is assigned to the nearest partition based upon some similarity parameter (such as Euclidean distance measure). After the completion of every successive pass, a data may switch partitions, thereby altering the values of the original partitions [8].
Various steps of the standard K-Means clustering algorithm is as follows [8]:

1. The number of clusters is first initialized and accordingly the initial cluster centers are randomly selected.
2. A new partition is then generated by assigning each data to the cluster that has the closest centroid.

3. When all objects have been assigned, the positions of the K centroids are recalculated.
4. Steps 2 and 3 are repeated until the centroids no longer move any cluster.

The flow chart of the k-means algorithm that means how the k-means work out is given in Figure 1 [9].

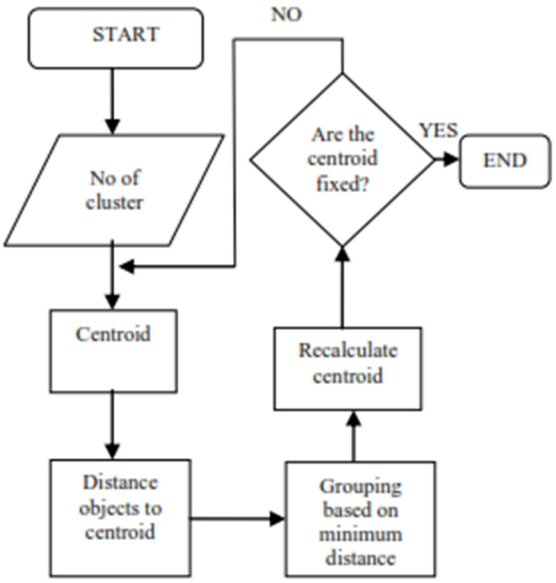

Fig. 1. Flowchart of K-Means Clustering [9]

3. GENETIC ALGORITHM

Genetic Algorithm (GA) is adaptive heuristic based on ideas of natural selection and genetics. Genetic algorithm is one of the most known categories of evolutionary algorithm. A GA works with a number of solutions which collectively is known as population in each iteration which is chosen randomly. These are adaptive heuristic search algorithms postulated on the evolutionary ideas of natural selection and genetic. The basic concept of these evolutionary algorithms is to stimulate process in natural system necessary for evolution. GA's are used for numerical and computational optimization and based on study the evolutionary aspects of models of social systems. The GA performs a balanced search on various nodes and there is a need to retain population diversity exploration so that any important information cannot be lost because there is a great need to focus on fit portions of the population.

The simplest form of genetic algorithm involves three types of operators: selection, crossover, and mutation [10].

Selection. This operator selects chromosomes in the population for reproduction. The fitter the chromosome, the more times it is likely to be selected to reproduce.

Crossover. This operator randomly chooses a locus and exchanges the subsequences before and after that locus between two chromosomes to create two offspring. For example, the strings 10000100 and 11111111 could be crossed over after the third locus in each to produce the two offspring 10011111 and 11100100. The crossover operator roughly mimics biological recombination between two single−chromosome (haploid) organisms.

Mutation. This operator randomly flips some of the bits in a chromosome. For example, the string 00000100 might be mutated in its second position to yield 01000100. Mutation can occur at each bit position in a string with some probability, usually very small (e.g., 0.001).

The process of genetic algroritma can be seen in Figure 2.

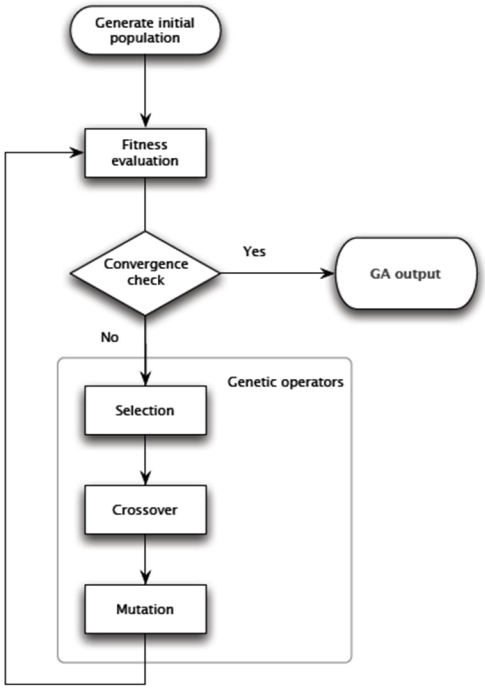

Fig. 2. Flowchart of Genetic Algorithm [11]

4. UCI MACHINE LEARNING REPOSITORY

The data used in the benchmark data are taken from the UCI Machine Learning Repository. UCI Machine Learning Repository is a collection of databases, domain theory, and data generator that is used by the community to study machine learning (machine learning), for the purposes of empirical analysis of machine learning algorithms. The dataset is available on the UCI Machine Learning Repository used by students, educators, and researchers around the world as the primary source of data sets in machine learning. The number of data sets available on the UCI Machine Learning Repository at the moment have amounted to 320 data sets can be used according to

the needs of the learning machine learning. The data set used is Iris Data Set that has the amount of data as much as 150 to 4 attributes.
Iris data set is a data set that is widely used in pattern recognition problems. Attribute information on Iris Data Set is composed-of: Sepal Length, Width Sepal, Petal Length and Width Petal. Iris Data Set has three classes, namely: Iris Setosa, Versicolour Iris, and Iris virginica.

5. DETERMINING A CLUSTER CENTRO-ID OF K-MEANS CLUSTERING USING GENETIC ALGORTIHM

5.1 Determining Number of Cluster

The number of cluster K in our study is 3(three) in accordance with the problems Iris Data Set that has 3 (three) classes, namely: Setos Iris, Iris versicolor, and Iris virginica.

5.2 Determining a Cluster Centroid

Because the number of clusters we have as many as three, then we are also use 3 cluster. We can use Genetic Algorithm in Determining the initial value of cluster centroid. Research on Iris Data Set using a 4 (four) attributes, namely: Sepal Length, Width Sepal, Petal Length and Width Petal. Length sepals have a value range 4.3 to 7.9, Sepal Width has a value range of 2 to 4.4, Petal Length has a value range of 1 to 6.9, and Petal Width has a value range of 0.1 to 2.5. Then the gene is generated in accordance with the range of each attribute there. because the value of the gene that are raised in the form of an integer, then the value range will be multiplied by 10.

5.3 Using Genetic Algortithm for Determining a Cluster Centroid

The process of the genetic algorithm are as follows:

1. Encode the parameter set according to the range of the four attribute(C1, C2, C3, C4) in the form of Bit Strings. The number of The String indicates the Number of Cluster.
   Bit strings are created with random assignment of 1 and 0 at different bit locations. We start with an initial population of five strings (Table 1). The Strings are 28 bits in length. The first 7 bits encode the C1, the second 7 bits encode the C2, the third 7 bits encode the C3 and the last 7 bits encode the C4.
   The value for the C1, C2, C3, and C4 are a decimal form of the string divided by 10.

Table 1: First Iteration Using Genetic Algorithm

String	C1	C2	C3	C4
0101110 0010110 0011111 0010010	4.6	2.2	3.1	1.8
0110100 0011110 0101111 0111001	5.2	3.0	4.7	5.7
1000111 0101001 1000000 0010111	7.1	4.1	6.4	2.3

2. Calculate the distance of the object in iris data set using euclidean distance.

$$dij = \sqrt{\sum_{k=1}^{p}(xik - xjk)^2} \quad (1)$$

Suppose we take dataset numbered 1 through 4 on Iris Dataset. The coordinate of each object can be shown in Table 2.

Table 2: Coordinate of Each Object

Object	Coordinate
Dataset 1	5.1 3.5 1.4 0.2
Dataset 2	4.9 3.0 1.4 0.2
Dataset 3	4.7 3.2 1.3 0.2
Dataset 4	4.6 3.1 1.5 0.2

Calculation of each object according to the Euclidean Distance is as follows.
Distance from Dataset 1 (5.1 3.5 1.4 0.2) to the Cluster 1 (4.6 2.2 3.1 1.8):

$$\sqrt{(5.1-4.6)^2 + (3.5-2.2)^2 + (1.4-3.1)^2 + (0.2-1.8)^2}$$
$$= 2.72$$

The distance of each object to the cluster centroid can be shown in Table 3.

Table 3: Coordinate of Each Object

Object	Distance to		
	Cluster 1	Cluster 2	Cluster 3
Dataset 1	2.72	6.43	5.81
Dataset 2	2.48	6.42	5.95
Dataset 3	4.33	6.48	4.58
Dataset 4	2.43	6.39	5.97

3. Grouping an object based on minimum distance
   According to Table 2, dataset 1, 2, 3, and 4 is grouped into the cluster 1.

4. We can compute the fitness of each string
   Fitness measurements in this study using Mean Square Error (MSE). The MSE calculate the error between the cluster of an object from the genetic algorithm process and an original from the Iris Dataset.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2 \quad (2)$$

$$Fitness = 1/\text{MSE} \quad (3)$$

5. We can go to the next generation or the next iteration if we want to get the better fitness than the fitness from generation 1.
   At this step we can go to the step of genetic algorithm, such as: selection, crossover, and mutation
6. Process will stop if we feel that we get the best fitness or the fitness value seems not to be changed to next generation.

6. RESULTS AND DISCUSSION

This research will be shown the results of performance assessment to initialize the cluster centroid using random numbers and by using genetic algorithms. Performance measurement is based on the Mean Square Error will be done using the method of calculation of Euclidean Distance. Performance value will be expressed in the form of Mean Square Error value which is the average value for the error based on tests using varying amount of training: 50, 75, and 100 that can be shown in Table 4, 5, and 6.

Table 4: Performance with Total Training = 50

Number of Testing	MSE from Genetic Algorithm	MSE from Random Number
1	1.68	1.78
2	0.78	0.81
3	0.63	0.72
4	0.77	0.76
5	1.33	1.21
6	1.08	0.99
7	0.99	1.07
8	0.87	1.02
9	1.21	1.4
10	1.32	1.21
Average of MSE	1.066	1.097

Table 5: Performance with Total Training = 75

Number of Testing	MSE from Genetic Algorithm	MSE from Random Number
1	1.27	1.17
2	1.09	0.91
3	0.58	0.65
4	0.91	0.89
5	1.03	1.22
6	0.88	0.89
7	0.91	1.02
8	0.65	0.64
9	0.67	0.7
10	0.88	0.81
Average of MSE	0.88	0.89

Table 6: Performance with Total Training = 100

Number of Testing	MSE from Genetic Algorithm	MSE from Random Number
1	1.2	1.02
2	1.2	0.92
3	0.67	0.66
4	0.88	0.88
5	0.65	0.9
6	0.78	1.02
7	0.92	0.99
8	0.67	0.68
9	0.58	0.61
10	0.77	0.79
Average of MSE	8.32	8.47

Based on Tables 4, 5, and 6 can be seen that the performance of the cluster centroid is initialized by using a genetic algorithm provides better results than using random numbers. This is demonstrated by a smaller MSE, if the initialization is done by using a genetic algorithm were compared using random number.

7. CONCLUSIONS

The conclusion that can be drawn from this study are as follows:

1. Genetic algorithms can be used in determining the initial value of the cluster centroid.
2. the performance of the cluster centroid is initialized by using a genetic algorithm provides better results than using random numbers.

REFERENCES

[1] Han, Jiawei, Kamber, Micheline and Pei, Jan. 2012. Data Mining: Concepts and Techniques. Morgan Kauffman: United States of America
[2] Leskovec, Jure, Rajaraman, Anand and Ullman, Jeffrey D. 2010. Mining of Massive Datasets. Madison Metropolitan School District: Palo Alto
[3] Zaki, Mohammed J., Meira Jr, Wagner. 2014. Data Mining and Analysis. Cambridge University Press: Cambridge
[4] Loohach, Richa and Garg, Kanwal. 2012. Efffect of Distance Functions on K-Means Clustering Algorithm. International Journal of Computer Application49(6): 7-9
[5] Liu, Dongxi, Elisa Bertino, and Yi, Xun. 2014. Privacy of Outsourced K-Means Clustering. Proceedings of the 9th ACM Symposium on Information, Computer, and Communication Security, CCS'14, pages: 123-134
[6] Huyen, Nguyen Thu, Uoc, LurongSy, Alday, Rosaly B. 2013. Genetic Algorithm for Solving Balanced Transportation Problem. International Journal of Innovative Technology and Exploring Engineering 3(4): pp. 23-27
[7] J. Mac Queen. 1967. Some methods for classification and analysis of multivariate observations. Fifth Berkeley Symposium on Mathematics, Statistics and Probability University of California Press: pp.281-297
[8] Mishra, Bikram Keshari and Nayak, Nihar Ranjan. 2012. Far Efficient K-Means Clustering Algorithm. Proceedings of the International Conference on Advances in Computing, Communications and Informatis, ICACCI '12, pages: 106-110
[9] Shovon, Md. Hedayatul Islam dan Haque, Mahfuza. 2012. An Approach of Improving Student's Academic Performance by Using K-Means Clustering Algorithm and Decision Tree. International Journal of Advanced Computer Science and Applications3(8): pp. 146-149
[10] Mitchel, Mielanie. 1999. An Introduction to Genetic Algorithm. MIT Press: Massachusets
[11] Scrucca, Luca. 2013. GA: A Package for Genetic Algorithm in R. Journal of Statistical Software53(4): pp. 1-37.

AUTHOR PROFILES:

Hartono received the Master degree in 2010 from the University of Putra Indonesia "YPTK" Padang, Indonesia in Computer Science and Bachelor Degree in 2008 from STMIK IBBI Medan, Indonesia in Computer Science. He is a lecturer at STMIK IBBI Medan. His current interests are in data mining and artificial intelligence. Nowadays, He Is a Student in a Doctoral Program in Computer Science at University of Sumatera Utara.

Erianto Ongko received the Master degree in 2015 from the University of Sumatera Utara, Indonesia in Computer Science and Bachelor Degree in 2012 from STMIK IBBI Medan, Indonesia in Computer Science. he is a designer and also Copilot at Top Coder Studio. His current interests are in design, data mining, and artificial intelligence.

Dahlan Abdullah completed a bachelor's degree in Informatics Engineering major at the Indonesia Islamic University in 1999. In 2014, he completed a master's degree in computer science at STMIK Eresha Jakarta. In 2015 was recorded as the doctoral students of Computer Science, University of North Sumatra, Medan. Actively doing research on Computer Networks, Database, Radio Net, Computer Applications, Robotics, Web Based Application, Management Information Systems and Computer Network Infrastructure.