Journal of Theoretical and Applied Information Technology
30th June 2012. Vol. 40 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

# ANALYSIS DATA OF STUDENT'S GPA AND TRAVELLING TIME TO CAMPUS USING CLUSTERING ALGORITHM AFFINITY PROPAGATION AND K-MEANS

**[1]S. JATMIKO, [2]R.REFIANTI, [3]A.B. MUTIARA, [4]R. WARYATI**

[1]Asst.Prof., Faculty of Computer Science and Information Technology, Gunadarma University, Indonesia

[2]Asst.Prof., Faculty of Computer Science and Information Technology, Gunadarma University, Indonesia

[3]Prof., Faculty of Computer Science and Information Technology, Gunadarma University, Indonesia

[4]Alumni, Faculty of Computer Science and Information Technology, Gunadarma University, Indonesia

E-mail: [1,2,3]{singgih,rina,amutiara}@staff.gunadarma.ac.id, [4]ihya25@gmail.com

**ABSTRACT**

In most of clustering algorithm, the cluster number and member are generated randomly, called common method. Consequently, the result may vary from the same clustering process. K-means algorithm is one example of these common methods. Affinity propagation, called new method, is developed to overcome this drawback by exchange the messages between data points to test the feasibility and accuracy of all data points to become exemplar and using it to select the cluster members. The aim of this research is to compare and analyse both methods by applying these method for clustering the student grade point average (GPA) and travel time to campus data. Both algorithms are implemented using Matlab 7.11.

**Keywords:** *Affinity Propagation, Clustering, Implementation, K-Means*

## 1. INTRODUCTION

Information needs from the available data cause the clustering algorithms continue to be developed to meet those needs. Various clustering algorithms including a new algorithm are developed based on previous algorithms. It aims to eliminate or reduce the drawback that occurred in the previous algorithm.

The problems that often arise in clustering are how to determine the number of clusters and how precise the results of the cluster is formed. Clustering algorithm generally begins with determining the number and member of cluster randomly. Consequently, different clusters are created for the same data and selection need to be performed to get the most optimal cluster. Some method can be applied to get the most appropriate cluster.

The well known K-Means algorithm [1] is a common strategy to solve various clustering problems. One of the drawback is the number of clusters should have be produced must be determine before clustering process. Consequently, the number of cluster produced can be less or more than the real number of clusters, by practically merging or separating the actual clusters. Therefore, the clustering process using K-Means usually is repeated to obtain the rationally number of cluster. In addition, the K-Means algorithm relies significantly on an initial choice of cluster centers. Instead of based on the true groupings inherent within the data, this choice may be random or influenced by previous clustering of similar data.

A large range of clustering algorithms are practically used, however these are usually customised for a specific domain. A general solution to the clustering problem is still needed. The new clustering algorithm, called Affinity Propagation, is offered by Brendan Frey and Delbert Dueck [2]. This method simultaneously considers all data points as potential exemplars and by employing a message passing procedure, the exemplars are "elected" by the other data points. The number of clusters is determined intuitively based on the data.

The aim of this research are:

- Analyse two clustering algorithms, Affinity Propagation and K-Means and relatively compare these algorithm based on the

Journal of Theoretical and Applied Information Technology
30th June 2012. Vol. 40 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

clustering process and results and relatively compare the clusters results

- Obtain the relationship between student's GPA and travelling time, by implementing both clustering algorithms

## 2. AFFINITY PROPAGATION ALGORITHM

Based on [1], affinity propagation is known in computer science as a message-passing algorithm, and advised that the algorithm can be understood by obtaining an anthropomorphic viewpoint. Affinity propagation is clustering algorithm that offers an interesting stage in the process of clustering the data**,** by implement ting message-passing algorithm in which the data clusters are formed based on the messages sent and received between each data point. Message will be sent by each data point to other data points to determine how well a data point as the cluster center (exemplar) and how well a data point is to become a member of a cluster [4] [5].

Fundamentally, the algorithm works on three matrices, which are a similarity (*s*), a responsibility (*r*), and availability (*a*) matrix. Results are contained in a criterion (*c*) matrix. These matrices are updated iteratively by four equations, where *i* and *k* refer, respectively, to the rows and columns of the associated matrix. Stage of affinity propagation clustering is shown in the following stages:

(1). At the first stage, similarity of the data points to all other data points are calculated. Similarity is the negative value of the distance based on particular distance calculation method. The method used in this research is, the simple common distance calculation, Euclidean distance. A certain value is used for similarity data point, called preference. The value of preference used is usually the median or the minimum of the entire similarity data. The number of clusters formed will determined by preference. Smaller preference will establish a smaller number of clusters and the use of larger preference will result in more clusters formed.

$$s(k,k) = p \forall k \in \{1, \cdots, N\}$$

(2). Initialization process is the initialization for availability. Initial values of availability is zero.

$$\forall i,k: a(i,k) = 0$$

(3). Calculation is performed to determine responsibility of data point to all other data points. The data point responsibility contains a message which is sent by the data point to exemplar candidate. Responsibility denoted by *r(i,k)* which represents the message sent by the data point i to exemplar candidate k about how well the data point k to be a exemplar candidate for data point i.

$\forall i,k$:

$$r(i,k) = s(i,k') - max_{k':k' \neq k}[s(i,k') + a(i,k')]$$

(4). Determining the availability of data point to all other data point and to the data point itself. Availability contains the message sent by the candidate exemplar to the data points. Availability is denoted by a (i, k) where a (i, k) represents the message sent by the candidate exemplar k to the data point i of how precise the data point k to be a candidate exemplar for data point i. Availability formulation is as follows:

$$\forall i,k: \\ a(i,k) = \sum_{i':i' \neq i} max[0, r(i',k)], for\ k = i$$

$$\forall i,k: \\ a(i,k) = \\ min\big[0, r(k,k) + \\ \sum_{i':i' \notin \{i,k\}} max[0, r(i,k)]\big], \ for\ k \neq i$$

(5). The step (3) is repeated if availability have not yet convergence or the net similarity is still undergoing change. Net similarity is the sum of similarity of data points and similarity of exemplar.

(6). The exemplar is determined by selecting the greatest addition of availability and responsibility where the data point k is set as the exemplar for data point i.

$$c(i,k) = r(i,k) + a(i,k)$$

## 3. K-MEANS ALGORITHM

K-means [6] is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. K-means clustering is non-hierarchy cluster algorithm that divides the data into one or more clusters based on similarity characteristics. The data has the same characteristics will be grouped in the same cluster.

Journal of Theoretical and Applied Information Technology
30th June 2012. Vol. 40 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Clustering using k-means algorithm is generally performed with the following algorithm:

(1). The number of clusters to be formed id determined. The number of clusters formed is not always equal to the number of clusters to be formed.

(2). Each data is allocated into one cluster to form clusters randomly. The desired number of clusters is based on the previous stage.

(3). The average distance, called centroid is calculated for all the data contained in the same cluster.

(4). The similarity distance between the centroid is calculated. Euclidean distance is used for similarity.

(5). Data is allocated into a cluster based on the closest similarity between centroid and data.

(6). Step 3 is called if the displacement is still happening or if there is a change value of centroid.

## 4. DESIGN AND IMPLEMENTATION

Two UML diagrams is used to describe the application and implemented in MATLAB 7.11. Both of diagrams is usecase and activity diagram. Usecase diagram describes the interactions that occur between user and the functional of the application. Meanwhile, Activity diagram describe the flow of activities that occur. System is designed with two clustering functionalities, K-Means and Affinity Propagation clustering, as can be seen in figure 1 in usecase diagram form.

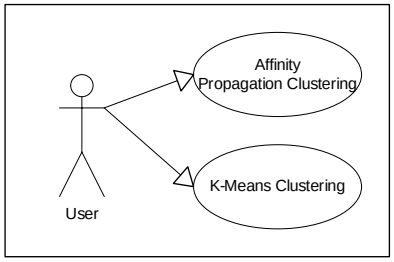


Figure 1. Usecase Diagram

The activities that occur in the application is shown in activity diagram in the figure 2. In Affinity propagation algorithm, preference input is needed before performing clustering. In K-Means, number of cluster is needed before performing the algorithm. The rest of activities are the same between Affinity Propagation and K-Means, including data loading and output.

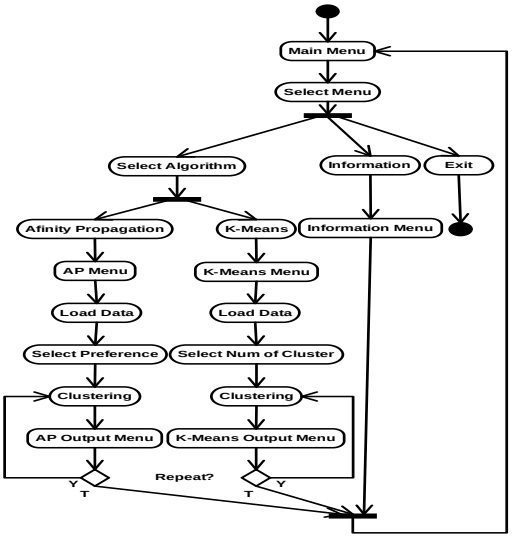


Figure 2. Activity Diagram

## 5. TESTING

The student grade point average (GPA) and travel time to campus data is used for testing clustering algorithms, as can be seen in table 1 below.

Table 1. GPA and Travel Time Data

| No | Data | | No | Data | |
|---|---|---|---|---|---|
| | GPA | Time minute | | GPA | Time minute |
| 1 | 2.56 | 40 | 26 | 3.07 | 25 |
| 2 | 3.61 | 50 | 27 | 2.55 | 75 |
| 3 | 3.00 | 45 | 28 | 2.97 | 35 |
| 4 | 3.56 | 40 | 29 | 2.31 | 60 |
| 5 | 3.38 | 60 | 30 | 2.47 | 45 |
| 6 | 2.78 | 75 | 31 | 2.81 | 40 |
| 7 | 2.86 | 25 | 32 | 2.77 | 75 |
| 8 | 3.47 | 30 | 33 | 3.21 | 40 |
| 9 | 3.38 | 75 | 34 | 2.82 | 30 |
| 10 | 2.75 | 30 | 35 | 3.38 | 50 |
| 11 | 2.93 | 15 | 36 | 2.60 | 30 |
| 12 | 2.74 | 45 | 37 | 3.28 | 75 |
| 13 | 3.13 | 60 | 38 | 2.68 | 65 |
| 14 | 3.27 | 45 | 39 | 3.19 | 90 |
| 15 | 3.81 | 40 | 40 | 2.98 | 55 |
| 16 | 3.21 | 60 | 41 | 3.46 | 60 |
| 17 | 3.34 | 60 | 42 | 2.83 | 30 |
| 18 | 3.18 | 80 | 43 | 3.33 | 20 |
| 19 | 2.97 | 90 | 44 | 2.53 | 20 |
| 20 | 2.76 | 75 | 45 | 3.41 | 80 |
| 21 | 3.45 | 65 | 46 | 2.69 | 75 |
| 22 | 3.62 | 45 | 47 | 3.07 | 70 |
| 23 | 3.27 | 60 | 48 | 2.95 | 25 |
| 24 | 3.12 | 15 | 49 | 3.17 | 30 |
| 25 | 3.15 | 75 | 50 | 2.67 | 45 |

Journal of Theoretical and Applied Information Technology
30th June 2012. Vol. 40 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

## 6. RESULTS AND COMPARISON

The clustering results of affinity propagation algorithm using the minimum similarity as the preference indicated in figure 3

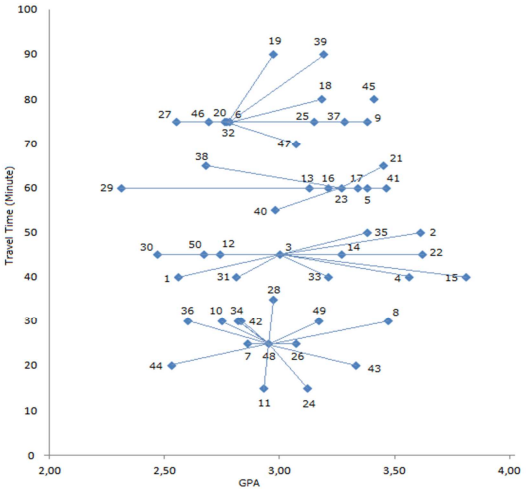


Figure 3. The Clustering Results of Affinity Propagation Algorithm

Clustering results using k-means algorithm with 4 as number of the initial cluster is shown in the following figures.

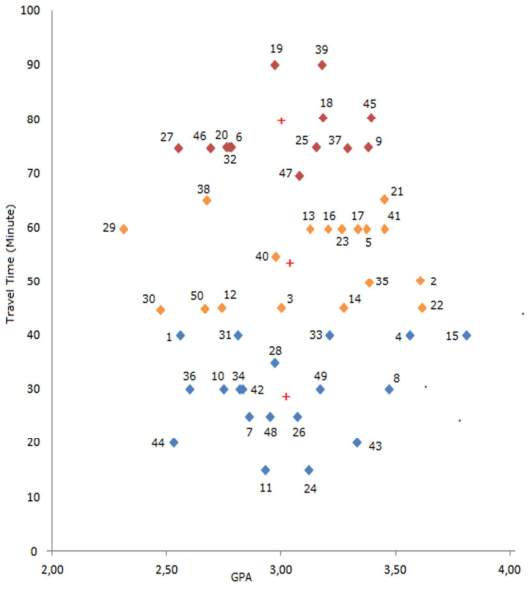


Figure 4. The 1st Clustering Results of K-Means Algorithm

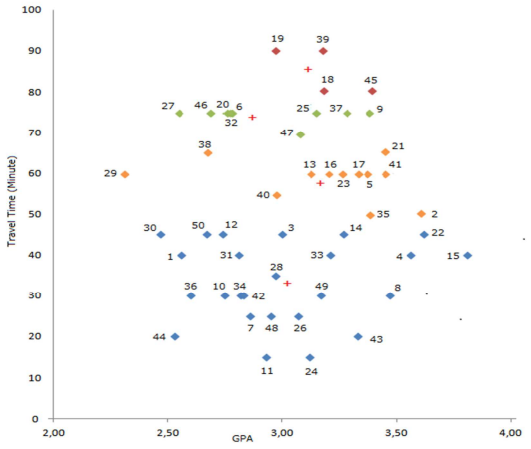


Figure 5. The 2nd Clustering Results of K-Means Algorithm

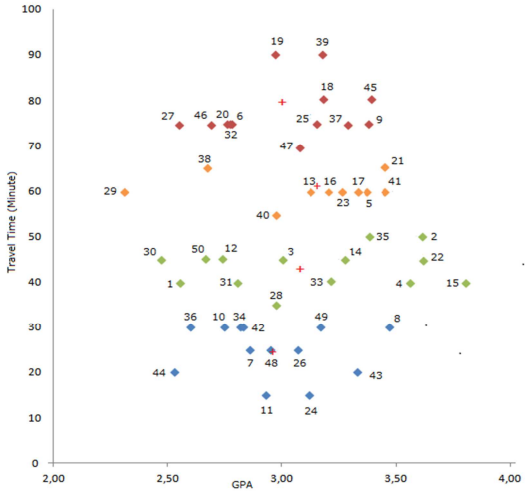


Figure 6. The 3rd Clustering Results of K-Means Algorithm

After clustering with the same data on both algorithms, the following results is obtained:

(1). Using the affinity propagation algorithm with minimum of similarity as preference formed 4 clusters are the same and no change exemplars and cluster members formed.

(2). In 3 trials using k-means algorithm with 4 as initial clusters provide different clusters. In the first experiment formed 3 clusters, the second and third experiment formed 4 clusters.

(3). The results of the cluster by k-means algorithm are sometimes not fixed. Both the number of clusters as well as members of the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

cluster formed. The number of clusters that formed in the cluster does not always correspond with the initial clusters is determined.

(4). Based on the similarity members of each cluster, the similarity of the affinity propagation clusters result with the first, second and third K-means clusters result are 36%, 72% and 98% respectively. On average, the similarity between both algorithms is 69%. Based on the results, comparing particularly affinity propagation clustering with K-means clustering can be varying from low to high similarity.

(5). More time is needed to perform the K-means clustering compare with the time to perform affinity propagation clustering.

## 7. CONCLUSION

The same result is provided by two trials using affinity propagation algorithm, with the minimum of similarity as preference. The different results can be provided either in number of result cluster or members of particular cluster. Tests are also carried out using small amounts of data and. Only two variables used in the implementation can be increased to more than two variables.

Based on both clusters formed also concluded that travel time does not affect the GPA. The resulted clusters of both algorithms can be varying from low to high similarity, but on average, the similarity is 69%. Compare with K-means, affinity propagation clustering is faster.

The research need to be expanded in some direction, including exploring a method to get the best number of clusters in K-Means algorithms, implementing with the large number of data and variables. Therefore, both algorithms can be compared comprehensively.

## REFRENCES:

[1] Jain, A. K., M. N. Murty, and P. J. Flynn. "Data Clustering: A Review." *ACM Computing Surveys* volume 31, issue #3, 1999, pp. 265-96.

[2] Frey, Brendan J., and Delbert Dueck. "Clustering by Passing Messages between Data Points." *Science* 315, 2007, pp. 972-76.

[3] Mézard, M. Where are the exemplars? *Science, 315*, 2007, pp. 949-951

[4] Dueck, D., Affinity Propagation: Clustering Data by Passing Messages. PhD-Thesis. Toronto : University of Toronto, 2009.

[5] Frey, B. J., & Dueck, D., Clustering by passing messages between data points. Science Vol 315, pp. 2007, 972-976.

[6] J. B. MacQueen, "Some Methods for classification and Analysis of Multivariate Observations, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability"*, Berkeley, University of California Press, 1967, pp. 281-297.