Implementation and Comparison of Affinity Propagation Algorithm and K-Means Clustering Data On Student GPA and Time Based Road to Campus

Prof. Dr.rer.nat. Dr.rer.nat. Achmad Benny Mutiara ¹ , Ria Waryati ²
Informatics Engineering, Faculty of Industrial Technology
Gunadarma University
Jl. Margonda Raya No. 100 Pondok Cina, Depok 16424, Indonesia
Email : amutiara@gmail.com ¹ , iyha25@gmail.com ²
¹) Dosen Teknik Informatika Universitas Gunadarma
²) Mahasiswa Teknik Informatika Universitas Gunadarma

Abstract

Determination of exact number of clusters and clustering results are the problems commonly experienced in the process of clustering. Existing clustering algorithms generally perform the process of determining the number and cluster members randomly, which result are not constant so it is necessary to determine the most appropriate cluster from the formed cluster.

Affinity propagation is a clustering algorithm that offers a new method, which is through the exchange of messages between data points to test the feasibility and accuracy of all data points to become the exemplar and the cluster members are selected based on the exemplar. While k-means clustering algorithm which is a data clustering process begins with determining the number and cluster members randomly. Authors performed the implementation and comparison of two clustering algorithms namely affinity propagation and k-means to determine the clustering algorithm that gives more precise results. Both algorithms are implemented using matlab 7.11 and tested against the data based on the value of student grade point average (GPA) and travel time to campus students.

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

1. Introduction

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

The problems that often arise in clustering are the determination of the number of clusters and the most precise results of the cluster is formed. Clustering algorithm generally begins with determining the number and cluster members randomly and resulting not fixed cluster and needs to determining the most optimal cluster from the cluster formed. It can improve the accuracy in determining the most appropriate cluster but require relatively longer time.

2. Affinity Propagation Algorithm

Affinity propagation is clustering algorithm that offers an interesting stage in the process of clustering the data, by implementation of message-passing algorithm in which the data clusters are formed based on the messages sent and received between each data point. Each data point will send a message to other data points on how well a data point to be the cluster center (exemplar) and how well a data point is to become a member of a cluster.

Stage of affinity propagation clustering is shown in the following stages:

1. Calculation similarity of the data points to all other data points. Similarity used is the negative value of the distance based on distance methods are used. At this writing the formulas used are euclidean distance. Similarity data point itself is called a preference and uses a certain value. Preference used is usually the median or the minimum of the entire similarity data. Preference will determine the number of clusters formed. Smaller preference will establish a number of smaller clusters. And the use of larger preference will result in more clusters are formed.

\[ s(k, k) = p \ \forall \ k \in \{1, ..., N\} \]
2. Initialization process is the initialization for availability. Initial values of availability is zero.
   $$\forall i, k : a(i, k) = 0$$

3. Perform calculations to determine responsibility of data point to all other data points. The responsibility contains a message which is sent by the data point to candidate exemplar. Responsibility denoted by $$r(i, k)$$ where $$r(i, k)$$ represents the message sent by the data point $$i$$ to candidate exemplar $$k$$ about how well the data point $$k$$ to be a candidate exemplar 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 $$k = i$$

5. Return to step 3 if have not convergence, if the net similarity is still undergoing change, then do step 3 again. Net similarity is the sum of similarity of data points and similarity of exemplar.

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

3. **K-Means Algorithm**

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.

Clustering using k-means algorithm is generally performed by the algorithm as follows:

1. Determining the number of clusters to be formed. The number of clusters formed is not always equal to the number of clusters to be formed.
2. Allocate each data into one cluster to form clusters the desired number of clusters in the previous stages randomly.
3. Calculate the average distance of all the data contained in the same cluster or called the centroid.
4. Calculate the similarity distance between the centroid is formed. The similarity used Euclidean distance.
5. Data is allocated into a cluster with the closest similarity between centroid and data.
6. Back to step 3 if the displacement is still happening or if there is a change value of centroid.

4. **Design and Implementation**

Program design built through 2 UML diagrams, usecase diagram 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. The draft application is implemented in MATLAB 7.11.

![Use Case Diagram](image-url)

**Figure 4-1 Use Case Diagram**

Use Case diagram in the figure describes the interactions that occur between user and the functional of the application. There are two cases, clustering using affinity propagation algorithm, and clustering using k-means algorithm.
Activity diagrams in the figure illustrate the activities that occur in the design application.

5. Testing Data
We use the data of student, shown in the following table:

**Table 5-1 Data of Student**

<table>
<thead>
<tr>
<th>No</th>
<th>GPA</th>
<th>Time (minute)</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>3.45</td>
<td>65</td>
</tr>
<tr>
<td>22</td>
<td>3.62</td>
<td>45</td>
</tr>
<tr>
<td>23</td>
<td>3.27</td>
<td>60</td>
</tr>
<tr>
<td>24</td>
<td>3.12</td>
<td>15</td>
</tr>
<tr>
<td>25</td>
<td>3.15</td>
<td>75</td>
</tr>
</tbody>
</table>

6. Results and Comparison
Afterwards performed clustering using affinity propagation algorithm and k-means. The clustering results of affinity propagation algorithm using the minimum similarity as the preference indicated in the following figure:

**Figure 6-1 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 figure:

**Figure 6-2 The Clustering Results of K-Means Algorithm**
7. **Comparison of Results**

After clustering with the same data on both algorithms can be obtained the following results:

1. Using the affinity propagation algorithm with minimum of similarity as preference formed 4 clusters are the same and no change exemplar 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 is sometimes not fixed. Both the number of clusters as well as members of the cluster formed. The number of clusters that formed in the cluster does not always correspond with the initial clusters is determined.
4. The results of the cluster is formed using these two algorithms show different results.

8. **Conclusion**

Trials using affinity propagation algorithm, using the minimum of similarity as preference provide the same results after performing 2 trials. The results 2 times the experiment of clustering using k-means algorithm with initial number of clusters of 4 clusters gives the clustering results are not always fixed by the number of clusters formed is not always consistent with the desired number of initial clusters. For the same number of clusters and centroid of each cluster member that is formed is impermanent. Clusters formed also concluded that travel time does not affect the GPA.

9. **Suggestion**

Affinity propagation algorithm can be developed with implementation in various fields such as the internet network or image processing. Tests carried out using small amounts of data, so it can be tested on large amounts of data. Variables used in the implementation can be increased to more than two variables.

**Bibliography**

