Performance Analysis Similarity Matrix, Responsibility Matrix, Availability Matrix, Criterion Matrix of Affinity Propagation

Hanna Willa Dhany
Faculty of Science and Technology, Universitas Pembangunan Pancabudi Medan, Indonesia
hdhany@dosen.pancabudi.ac.id

Abstract. Each process in classifying data into several clusters or grouping so that the data in one cluster has a maximum similarity level and between clusters has a minimum similarity is called clustering. Clustering is divided into 2 approaches in its development, namely the partitioning and hierarchical approach to clustering[1]. The Water Quality Status dataset has 8 attributes, 4 classes and 120 instances, Class distribution is good condition (30 instances), lightly polluted (30 instances), medium polluted (30 instances) and heavily polluted (30 instances). 70% of the data will be used as training data and 30% of the data will be used as randomized test data. The simplify the process of completing the performance calculation of the clustering model, the research implementation was carried out using the MATLAB function. That iteration is carried out with the number of clusters generated from 100 to 2,500 iterations with the results of the number of clusters as many as 10 clusters. In the experiment, iteration amounted to 5000 and there was a change in the results of the number of clusters by 9 clusters. After re-testing using the number of iterations of 10,000-50,000 iterations, but the number of clusters produced did not change anything at all. So that the conclusion in testing the AP method produces the most optimal number of clusters of 10 clusters.

1. Introduction
Each process in classifying data into several clusters or grouping so that the data in one cluster has a maximum similarity level and between clusters has a minimum similarity is called clustering. Clustering is divided into 2 approaches in its development, namely the partitioning and hierarchical approach to clustering[1].

[2] in their research tried to determine the Preference value based on the normalization of the total similarity (similarity) of each data point, this value is used for all Preference values. Based on these results, there is an increase in the results of clustering when compared to the standard Affinity Propagation algorithm, which is an increase in accuracy and the value of the Silhouette Coefficient. According [3] clustering is grouping objects (data) based only on the information contained in these objects and the relationships between these objects. Data grouping is usually done based on the similarity of values between data [4].

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd
AP is an algorithm that can process clusters created by Brendan J. Frey and Delbert Duek in 2007. The AP algorithm at a data point is seen in a network node, which will be processed when sending messages which are carried out by all data points in a repeated manner, thus forming a good exemplar set [5].

2. Research Methods

This study will examine performance of the AP algorithm in relation to the method of determining the Preference value. The determination of the Preference value will affect performance of the AP algorithm. Performance measurement is done by comparing the value of the Silhouettes Coefficient (Rousseeuw, 1986) generated from the standard Affinity Propagation algorithm, which determines the Preference value based on the median value of all data contained in the Similarity Matrix table with the modified Affinity Propagation algorithm.

Silhouette Coefficient is used to validate a single data cluster, or even an entire cluster. This method is widely used to validate clusters that combine cohesion and separation values. The range of values for the Silhouette Coefficient is -1 to +1. The Silhouette Coefficient value is close to 1 indicating that the data is right in the cluster, if the Silhouette Coefficient value is 0 or close to 0 then the data position is on the border of the two clusters, for the dataset \( X = \{x_1, x_2, \ldots, x_n\}, x_i \in \mathbb{R} \), the value of the Silhouette Coefficient (Rousseeuw, 1986).

\[
s(i) = \frac{(b(xi) - a(xi))}{nas(a(xi), b(xi))}
\]

where, \( a(x_i) \) is the average distance from other data points contained in the same cluster (intra-cluster distance), \( b(x_i) \) is the average distance from other clusters (inter-cluster distance).

3. Result and Discussion

In the results carried out by researchers in implementing the AP algorithm by supporting the program. The purpose of using this AP method is to be able to analyze grouping data that is more optimal when doing clusters. Therefore it is necessary to do an analysis of the AP method.

### Table 1. Data Normalization Results

| No | TSS  | DO  | COD  | BOD  | Total phosphate | ... | Pij  |
|----|------|-----|------|------|-----------------|-----|------|
| 1  | -0.90| -0.45| -0.44| -0.35| -0.42           | ... | -1.04|
| 2  | -0.89| -0.08| -0.20| -0.31| -0.24           | ... | -1.01|
| 3  | -0.89| -0.16| -0.27| -0.33| -0.33           | ... | -1.00|
| 4  | -0.87| -0.38| -0.51| -0.43| -0.06           | ... | -1.01|
| 5  | -0.87| -0.31| -0.44| -0.35| -0.37           | ... | -1.03|
| 6  | -0.85| -0.38| -0.44| -0.35| -0.42           | ... | -1.03|
| 7  | -0.85| -0.45| -0.44| -0.32| 0.07            | ... | -1.01|
| 8  | -0.83| -0.16| 0.07 | -0.38| -0.33           | ... | -0.99|
| ... | ...  | ...  | ...  | ...  | ...             | ... | ...  |
| 11 | 0.0534722 | 0.125 | 0.057 | 0.014 | -0.78           | ... | 0.04 |

The calculation of the Z-Score normalization for the TSS attribute (mg / L) from data to -1 is as follows:
\[ z = \frac{x - \mu}{\sigma} ; \quad z = \frac{5.323}{56.39} = -0.90 \]

\( z \): standard score, \( x \): observed data, \( \mu \): mean per variable and \( \sigma \): standard deviation per variable. The result of the Z-score is data with mean = 0 and standard deviation = 1.

The Water Quality Status dataset has 8 attributes, 4 classes and 120 instances, Class distribution is good condition (30 instances), lightly polluted (30 instances), medium polluted (30 instances) and heavily polluted (30 instances). 70% of the data will be used as training data and 30% of the data will be used as randomized test data.

The simplify the process of completing the performance calculation of the clustering model, the research implementation was carried out using the MATLAB function. The following is the result of the Similarity Matrix value:

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|c|c|}
\hline
Data & 1 & 2 & 3 & 4 & 5 & \ldots & 117 \\
\hline
1 & -25.64 & -0.2295 & -0.1232 & -0.1476 & -0.0231 & \ldots & -11.4627 \\
\hline
2 & -0.2295 & -25.64 & -0.0199 & -0.2333 & -0.1298 & \ldots & -9.8816 \\
\hline
3 & -0.1232 & -0.0199 & -25.64 & -0.1894 & -0.0547 & \ldots & -10.1019 \\
\hline
4 & -0.1476 & -0.2333 & -0.1894 & -25.64 & -0.1127 & \ldots & -11.6263 \\
\hline
5 & -0.0231 & -0.1298 & -0.0547 & -0.1127 & -25.64 & \ldots & -10.904 \\
\hline
6 & -0.0075 & -0.1836 & -0.0883 & -0.1417 & -0.0078 & \ldots & -11.0236 \\
\hline
7 & -0.2444 & -0.2923 & -0.2748 & -0.0392 & -0.2149 & \ldots & -11.7387 \\
\hline
8 & -0.3606 & -0.0963 & -0.1218 & -0.4622 & -0.2883 & \ldots & -9.4595 \\
\hline
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
\hline
117 & -11.4627 & -9.8816 & -10.1019 & -11.6263 & -10.904 & \ldots & -25.64 \\
\hline
\end{tabular}
\caption{Similarity Matrix Water Quality Status}
\end{table}

The similarity value obtained, the similarity value taken is the negation of the sum of the squares and the distance calculation method used in this study is the Euclidean distance. The specific value used for data similarity is called preference. The preference value taken for data is 25.64. Therefore, the diagonal value in the similarity matrix above is filled with a value of 25.64.

The next step is to calculate the responsibility value for the entire data and the number of iterations of 50,000 iterations. The following are the results of the acquisition of the responsibility value in the form of a matrix:

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|c|c|}
\hline
Data & 1 & 2 & 3 & 4 & 5 & \ldots & 117 \\
\hline
1 & -25.19236 & 1,4916667 & 2,2298611 & 2,0604167 & 2,925 & \ldots & -11.0184 \\
\hline
2 & 0,4979167 & -25.33546 & 1,9534722 & 0,4715278 & 1,1902778 & \ldots & -9.5804 \\
\hline
3 & 1,1020833 & 0,1819444 & -25.35476 & 0,6423611 & 1,5777778 & \ldots & -9.82 \\
\hline
4 & 2,0201389 & 1,425 & 1,7298611 & -25.19816 & 2,2625 & \ldots & -11.1878 \\
\hline
5 & 1,9576389 & 1,2166667 & 1,7381944 & 1,3354167 & -25.33166 & \ldots & -10.599 \\
\hline
6 & 2,3118056 & 1,0888889 & 1,7506944 & 1,3798611 & 2,3097222 & \ldots & -10.6832 \\
\hline
\end{tabular}
\caption{Responsibility Matrix Water Quality Status}
\end{table}
The diagonal value of the Responsibility Matrix reflects the self-responsibility value of the k-th column and i-row data. To obtain the responsibility value, the similarity value in the k-th column data is reduced by the minimum value of the i-row in the similarity matrix. Here are the results of the available value in the form of a matrix:

| Data | 1     | 2     | 3     | 4     | 5     | ...   | 117   |
|------|-------|-------|-------|-------|-------|-------|-------|
| 1    | 5.2423611 | -20.99206 | -20.76776 | -21.04586 | -20.39746 | ...   | -21.07806 |
| 2    | -20.53316 | 4.0430556 | -20.72796 | -20.81706 | -20.14766 | ...   | -21.07806 |
| 3    | -20.62016 | -21.03926 | 6.4729167 | -20.84166 | -20.20346 | ...   | -21.07806 |
| 4    | -20.75236 | -20.98246 | -20.69576 | 0.4784722 | -20.30206 | ...   | -21.07806 |
| 5    | -20.74336 | -20.95246 | -20.69696 | -20.94146 | 2.6763889 | ...   | -21.07806 |
| 6    | -20.79436 | -20.93406 | -20.69876 | -20.94786 | -20.30886 | ...   | -21.07806 |
| 7    | -20.76436 | -21.03226 | -20.71916 | -21.25726 | -20.30866 | ...   | -21.07806 |
| 8    | -20.56356 | -21.14366 | -20.78756 | -20.74966 | -20.15066 | ...   | -21.07806 |
| ...  | ...    | ...    | ...    | ...    | ...    | ...   | ...   |
| 117  | -20.46146 | -20.77726 | -20.44666 | -20.74916 | -19.97626 | ...   | 1.5708333 |

Table 4. Availability Matrix

The availability value of the water quality data, the diagonal availability value of the water quality status data is obtained by looking at the responsibility matrix value in the column (k-k) and only adding up the positive responsibility values (k value> 0) without including the rows with the same responsibility value with the (k-th) column. The next step is to add the availability value with the responsibility value so that it can be formed into a criterion matrix. The following is the result of the criterion matrix value obtained from the water quality status data after 50,000 iterations:

| Data | 1     | 2     | 3     | 4     | 5     | ...   | 117   |
|------|-------|-------|-------|-------|-------|-------|-------|
| 1    | -20.46146 | -20.77726 | -20.44666 | -20.74916 | -19.97626 | ...   | -32.09646 |
| 2    | -20.46146 | -20.77726 | -20.44666 | -20.74916 | -19.97626 | ...   | -30.65846 |
| 3    | -20.46146 | -20.77726 | -20.44666 | -20.74916 | -19.97626 | ...   | -30.89806 |
| 4    | -20.46146 | -20.77726 | -20.44666 | -20.74916 | -19.97626 | ...   | -32.26586 |
| 5    | -20.46146 | -20.77726 | -20.44666 | -20.74916 | -19.97626 | ...   | -31.67706 |
The criterion value is obtained by adding the value in the availability matrix with the value in the responsiveness matrix. The highest value on the criterion matrix of each row (ith) is called an exemplar (center point). Exemplar is data that is selected from existing data to represent and from other data. The table below shows the exemplar value obtained from the application of the AP method.

| No. | Cluster | Exemplar |
|-----|---------|----------|
| 1   | C1      | 30       |
| 2   | C2      | 55       |
| 3   | C3      | 72       |
| 4   | C4      | 74       |
| 5   | C5      | 76       |
| 6   | C6      | 80       |
| 7   | C7      | 86       |
| 8   | C8      | 88       |
| 9   | C9      | 107      |
| 10  | C10     | 109      |

4. **Conclusion**

In conclusion, the results show that these two parameters greatly influence the final clustering of the AP algorithm. The Water Quality Status dataset has 8 attributes, 4 classes and 120 events, class distribution is in good condition (30 events), lightly polluted (30 events), moderately polluted (30 events) and heavily polluted (30 events). The preference value taken for the Water Quality Status data is 25.64. the value of the response matrix in the column (k-k) and only adds up the positive responsibility (value k> 0) without including the row whose responsiblility value is the same as the column (k-k). The criterion value is obtained by adding the value in the availability matrix with the value in the responsibility matrix. The highest value on the criterion matrix of each row (ith) is called an exemplar (center point). The resulting exemplar value contains 10 clusters, which is the highest value that has been selected from data that is represented from other data. So that iteration is carried out with the number of clusters generated from 100 to 2,500 iterations with the results of the number of clusters as many as 10 clusters. In the experiment, iteration amounted to 5000 and there was a change in the results of the number of clusters by 9 clusters. After re-testing using the number of iterations of 10,000-50,000 iterations, but the number of clusters produced did not change anything at all. So that the conclusion in testing the AP method produces the most optimal number of clusters of 10 clusters.
References
[1] Poteras, Cosmin Marian., Mihaescu, Marian Cristian., Mocanu, Mihai. 2014. An Optimized Version of the K-Means Clustering Algorithm. University of Craiova Faculty of Automation, Computers and Electronics Blvd. Decebal nr. 107, Craiova, Romania.
[2] Zhang, Kang., Xingsheng Gu. 2014. An Affinity Propagation Clustering Algorithm for Mixed Numeric and Categorical Datasets. Key Laboratory of Advanced Control and Optimization for Chemical Processes, Ministry of Education, East China University of Science and Technology, Shanghai 200237, China.
[3] Tan, Pang-Ning, et al. Introduction to data mining. Boston: Pearson Addison Wesley, 2006
[4] Xia, L., Mengqiao, H., Fang, G., & Peihong, X. (2008). An empirical study of online shopping customer satisfaction in China: a holistic perspective. International Journal of Retail & Distribution Management, 36, 919–928.
[5] Brendan J. Frey., Delbert Dueck. 2007. Clustering by Passing Messages Between Data Points. Science, 315(5814), 972–976.