Performance Improvement of Clustering Affinity Propagation Method using Principal Component Analysis

Jasael Simanullang¹, Muhammad Zarlis², Elviawaty Muisa Zamzami³
¹Graduate Program of Computer Science,  
²Faculty of Computer Science and Information Technology,  
³Universitas Sumatera Utara, Medan, Indonesia

Email : ¹jasael.simanullang@yahoo.co.id, ²m.zarlis@usu.ac.id, ³elvi_zamzami@usu.ac.id

Abstract. Affinity Propagation Method it is necessary to modify the algorithm by using Principal Component Analysis (PCA). PCA method is used to reduce the attributes or characteristics that are less influential on the data so that the most influential attributes are obtained to then be carried out the clustering process with Affinity Propagation. The comparison results of the PCA + AP grouping model have better performance than the conventional AP grouping model. This is justified because the number of iterations and clusters produced by the PCA + AP clustering model does not change and converges when there are 8 optimal cluster clusters. While the performance of conventional clustering models produces an optimal number of clusters from 14 clusters with a significant number of iterations. So it can be concluded that the PCA + AP grouping model is suitable for the Air Quality dataset because it produces an optimal number of clusters and iterations of 8 clusters. The comparison results of the PCA + AP grouping model have better performance than the conventional AP grouping model. This is justified because the number of iterations and clusters produced by the PCA + AP clustering model does not change and converges when the optimal number of clusters is 5 clusters. While the performance of conventional clustering models produces a suboptimal number of 10 clusters with a significant number of iterations. So it can be concluded that the PCA + AP grouping model is suitable for the Water Quality Status dataset because it produces an optimal number of clusters and 5 cluster repetitions.

1. Introduction

Clustering is part of the unsupervised learning method because it does not require defining the cluster first (Nisha and Kaur, P.J. 2015). In clustering the measurement of similarity between objects is done by measuring the distance for each pair of objects. This measurement can be done by the Euclidean Distance, Manhattan Distance and Minkowski Distance methods.

One clustering algorithm is the Affinity Propagation algorithm, this clustering algorithm was introduced by Brendan J. Frey and Delbert Dueck in 2007. Unlike other clustering algorithms such as K-Means which take random data points to be used as potential exemplar candidates, Affinity Propagation uses a different approach in determining exemplar, Affinity Propagation considers all data points as potential exemplar candidates (Frey BJ and Dueck, D. 2007).

PCA gives good results when applied to correlated attributes. In this study, PCA was applied in training and testing factors that greatly influenced the dataset. PCA will identify patterns in the data set, find similarities and differences between each factor.

Hussain et al. (2015) used the Principal Component Analysis (PCA) approach as a method of feature selection to reduce indicators related to prediction of survival rates of patients infected with breast cancer. The data used were from the SEER dataset of 684,394 patient medical records. With the proposed approach, the accuracy is 92%.

Principal Component Analysis (PCA) approach is expected to be able to simplify and eliminate factors that are less dominant or relevant to affect the data tested but have a large correlation to the formation of the tested data factors with a total proportion of expected variance of covariance of 60%. So this makes it easier for educational institutions to further improve the accuracy of the data being tested.
Weaknesses in the clustering method can affect performance in clustering data. Performance can be interpreted as the level of achievement of results. Therefore, based on previous studies, this research is proposed with the aim of increasing the performance of the Affinity Propagation method using Principal Components Analysis (PCA), it is hoped that this will be able to overcome the weaknesses in Affinity Propagation and result in improved performance in clustering the data used.

2. Research Methods

Affinity Propagation is an algorithm for the clustering process introduced by Brendan J. Frey and Delbert Dueck in 2007. In the Affinity Propagation algorithm all data points are seen as a node on the network, then the process of sending messages is carried out throughout all data points repeatedly, until they are formed a set of good exemplars (Frey et al. 2007). Exemplar is defined as the best data point to represent data.

The process of Affinity Propagation algorithm can be seen as the process of exchanging two types of messages between data points. These messages are "responsibility" r(i, k) and "availability" a(i, k) (Fujiwara et al. 2011), these two types of messages can determine which points are Exemplar and points that are become part of each Exemplar (Ding, L and Mighu, J. 2012). The process of sending messages that occur is seen as in the following image.

The Affinity Propagation Method is a method of clustering a message with real value between data points until high-quality copies obtain a suitable cluster. The Affinity Propagation Method assumes that all data points have the same opportunity to become copies (the central point). A copy is data selected from all data that represents itself and other data (Frey & Dueck, 2007).

Each similarity shows how well the data points with the corresponding k-index are copies for the i-data points. The goal is to minimize square error so that each similarity has a negative sum square error value. There are two types of messages exchanged between data points so that the message can be combined at any stage to determine which point is a copy. The first message is Responsibility (i, k), which sends the data-i point to the candidate copy k, measuring how well the dot-k is to be copied to the i-point The second message is Availability (i, k), which is sent from the candidate point-k copies to point-i, thus showing how feasible the point-i is to choose point-k as its copy (Frey and Dueck, 2007).

3. Problems Identification

Based on the above background in improving the clustering process of the Affinity Propagation method, it is necessary to modify the algorithm by using Principal Component Analysis (PCA). PCA method is used to reduce the attributes or characteristics that are less influential on the data so that the most influential attributes are obtained to then be carried out the clustering process with Affinity Propagation.

4. Result and Discussion

In this section, it describes the implementation of research with the support of Matlab version R2008a and Rapid Miner® version 5.3. The proposed method is expected to improve the performance of the Affinity Propagation method using Principal Component Analysis (PCA) so as to produce an optimal cluster when the clustering process is carried out. Therefore, the performance of the proposed method is compared with the conventional Affinity Propagation method to determine the good performance of the method by looking at the quality and strength of the clustering produced.

Table 1 Results of Cleaning Quality Water Status Data
| Actual Class | Good | Dangerous | Very not. | Half | No Data | No Healthy |
|--------------|------|-----------|-----------|------|---------|------------|
| Good         | 144  | 0         | 0         | 4    | 0       | 0          |
| Dangerous    | 0    | 6         | 0         | 0    | 0       | 0          |
| Very not.    | 0    | 0         | 1         | 0    | 0       | 0          |
| Healthy      | 7    | 0         | 0         | 38   | 0       | 1          |
| Half         | 0    | 0         | 0         | 0    | 0       | 0          |
| No, there isn’t | 1 | 0       | 0       | 2    | 0     | 12         |

**Table 2** Results of Air Quality Cleaning Data in Pekanbaru City

| No  | PM10 | SO₂ | CO | O₃ | NO₂ | Category |
|-----|------|-----|----|----|-----|----------|
| 1   | 47   | 51  | 8  | 67 | 2   | Half     |
| 2   | 48   | 51  | 9  | 37 | 2   | Half     |
| 3   | 37   | 51  | 9  | 26 | 2   | Half     |
| 4   | 24   | 50  | 2  | 51 | 1   | Half     |
| 5   | 24   | 50  | 2  | 51 | 1   | Half     |
| 6   | 25   | 50  | 3  | 36 | 1   | Good     |
| 7   | 18   | 50  | 3  | 53 | 2   | Half     |
| 8   | 15   | 50  | 6  | 61 | 1   | Half     |
| 9   | 20   | 50  | 4  | 36 | 1   | Good     |
| 10  | 27   | 50  | 7  | 36 | 2   | Good     |
| ... | ...  | ... | ...| ...| ...  | ...      |
| 1080| 37   | 11  | 7  | 23 | 7   | Good     |

4.1. **Results of PCA Analysis**

The following is the result of the decomposition of the Water Quality Status dataset and the Air Quality dataset using Matlab:

The Eigen value is obtained by multiplying the variance proportion value with the total variance covariance of the attributes. Can be seen as follows:

Eigen value = Proportion of Variants (%) x Covariance Variance

= 27.10 x 8

= 216.80

on in the same way also applies to each PC

Eigen value decomposition results from the Pekanbaru City Air Quality dataset using the Eigen Value Decomposition (EVD) equation. The Eigen value is obtained by multiplying the variance proportion value with the total variance covariance of the attributes. Can be seen as follows:

Eigen value = Proportion of Variants (%) x Covariance Variance

= 56.90 x 5

= 284.50

on in the same way also applies to each PC
In this study, the number of principal components chosen is the maximum proportion of variance covariance that is able to explain the variance of covariance from the original attribute. The proportion of variance covariance taken for the Water Quality Status dataset is the proportion of cumulative variance of 50.90% obtained from the sum of the values of the proportion of variance from the 1st principal component to the 2nd principal component so that a number of 2 priority components is obtained as follows:

| PC | Nilai Eigen | Proporsi Varian (%) | Cumulative |
|----|-------------|---------------------|------------|
| 1  | 216.8       | 27.1                | 27.1       |
| 2  | 191.2       | 23.9                | 50.9       |

*Variance Threshold = 50.90%*

The proportion of covariance variance taken for the Air Quality dataset was 56.90% obtained from the value of the variance proportion in the 1st principal component as follows:

| PC | Nilai Eigen | Proporsi Varian (%) | Cumulative |
|----|-------------|---------------------|------------|
| 1  | 284.5       | 56.9                | 56.9       |

*Variance Threshold = 56.90%*

The performance of the conventional Affinity-Propagation model for Water Quality Status data, then an analysis of performance based on the number of iterations so as to produce the optimal number of clusters during the clustering process is carried out. Here is an analysis and comparison chart of the conventional Affinity-Propagation model for the Water Quality Status dataset, seen in the following picture:

![Graph of Iterasi & Cluster](chart1)

*Figure 1 Graph of Conventional Affinity-Propagation Analysis (Dataset Water Quality Status)*

The performance of the conventional Affinity-Propagation model for Pekanbaru City Air Quality data, a performance analysis is performed based on the number of iterations that occur so as to produce the optimal number of clusters during the clustering process. Here is an analysis and comparison chart of the conventional Affinity-Propagation model for the City Air Quality dataset Pekanbaru, can be seen in the following picture:
It was explained that iterations had been carried out with a range and number of clusters produced sequentially starting from 100, 250, 1,000, 2,000 iterations which resulted in the same number of clusters namely 14 clusters. However, when iterated 500 and 2,500 iterations, the number of clusters produced was 15 cluster each. After iterating as many as 5,000 to 50,000 iterations the number of clusters did not change ie as many as 14 clusters and the value of copies obtained did not also change so by using the conventional Affinity-Propagation model for the Air Quality dataset Kota Pekanbaru has produced an optimal number of clusters of 14 clusters.

4.2. Test Results Model PCA + Affinity-Propagation (PCA + AP)

To see the performance of the PCA + AP model for the Water Quality Status dataset, a performance analysis is performed based on the number of iterations that occur so as to produce the optimal number of clusters during the clustering process. Here is an analysis chart and a comparison of the PCA + AP model against the Water Quality Status data set, it can be seen in the following picture:

The results of the clustering model analysis use iteration with a range and number of clusters starting from 100 iterations which results in a cluster of 64 clusters. This will be followed by 250 iterations which will produce as many as 43 cluster clusters. However, when iterated 500, 1,000, 2,000, 2,500, 5,000, 10,000, 25,000 and 50,000 iterations there was a significant change in the number of clusters and the number of clusters remained unchanged is as many as 5 clusters and the copies obtained also did not change, thus using the PCA + Affinity-Propagation model for the Water Quality Status dataset has resulted in an optimal number of clusters of 5 clusters.

4.3. Air Quality Data Clustering Results (PCA + AP)

The comparison chart of the performance of the PCA + AP clustering model against the Pekanbaru Water Quality Status and Air Quality dataset, can be seen in the following figure:
The results of the PCA + Affinity Propagation clustering process performance on the Water Quality Status dataset obtained from the analysis of 500 iterations, 1,000, 2,000, 2,500, 5,000, 10,000, 25,000 and 50,000, the number of fixed and unchanged clusters of 5 clusters and the value of copies obtained also unchanged, thus the optimal number of clusters is 5 clusters. While the results of the PCA + Affinity Propagation model clustering process performance on the Air Quality dataset obtained by the analysis of as many iterations 100, 250, 500, 1,000, 2,000, 2,500, 5,000, 10,000, 25,000 and 50,000, the number of clusters is also constant and has not changed, as many as 8 clusters so that using the PCA + Affinity-Propagation model for the Air Quality dataset produces an optimal number of clusters 8 clusters.

5. Conclusion

The comparison results of the PCA + AP clustering model have better performance than the conventional AP clustering model. This is justified because the number of iterations and clusters produced by the PCA + AP clustering model has not changed and is convergent when there are 8 optimal cluster clusters. While the performance of the conventional clustering model results in a suboptimal number of clusters of 14 clusters with a significant number of iterations. So it can be concluded that the PCA + AP clustering model is suitable for the Air Quality dataset because it produces an optimal number of clusters and iterations of 8 clusters.

The comparison results of the PCA + AP clustering model have better performance than the conventional AP clustering model. This is justified because the number of iterations and clusters produced by the PCA + AP clustering model has not changed and is convergent when the optimal cluster number is 5 clusters. While the performance of the conventional clustering model results in a suboptimal number of clusters of 10 clusters with a significant number of iterations. So it can be concluded that the PCA + AP clustering model is suitable for the Water Quality Status dataset because it produces an optimal number of clusters and iterations of 5 clusters.

References

[1]. Frey, B.J., Dueck, D. 2007. "Clustering by passing messages between data points". Science 315 (5814): 972–976.
[2]. Guha, S., Rastogi, R. & Shim, K. 1998. Cure: An efficient clustering algorithm for large databases. In Proceedings of ACMSIGMOD International Conference on Management of Data, pp.73 – 84.
[3]. Hussain, S., Quazilbash. N.Z., Bai. S. & Khoja, S. 2015. Reduction of Variables for Predicting Breast Cancer Survivability Using Principal Component Analysis.International Conference on Computer-Based Medical Systems, pp. 131-134.
[4]. Johnson, W.A. & Wichern, D.W. 2007. *Applied Multivariate Statistical Analysis*. 6th Edition. Pearson Prentice Hall: New Jersey.

[5]. Jolliffe, I.T. 2002. *Principal Component Analysis*. 2nd Edition. Springer-Verlag: New York.

[6]. Kavitha, R. & Kannan, E. 2016. An Efficient Framework for Heart Disease Classification using Feature Extraction and Feature Selection Technique in Data Mining. *International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS)*, pp. 1-5.

[7]. Kotu, V. & Deshpande, B. 2015. *Predictive Analytics and Data Mining*. Morgan Kaufmann Publisher: San Francisco.

[8]. Refianti, R., Mutiara, A.B., Juarna, A., Ikhsan, S.N. 2012. Analysis and Implementation of Algorithm Clustering Affinity Propagation and K-Means at Data Student Based on GPA an Duration of Bachelor-Thesis Completion. Journal of Theoretical and Applied Information Technology. Vol. 35 No. 1. pp. 69-76

[9]. Steinbach, M., Karypis, G. & Kumar, V. 2000. A comparison of document clustering techniques. *KDD Workshop on Text Mining*, pp. 525.

[10]. Thavikulwat, P. 2008. Affinity Propagation: A Clustering Algorithm for Computer-Assisted Business Simulations and Experiental Exercises. Developments in Business Simulations and Experiental Learning. Vol. 35. pp. 220-224.