

Modified affinity propagation

Heru Gunawan^{1,4}, Saib Suwilo^{2,5}, Zakarias Situmorang^{3,6}

¹Departement of Computer Science, Universitas Sumatera Utara, Medan, Indonesia

²Department of Mathematics, Universitas Sumatera Utara, Medan, Indonesia

³ Universitas Katolik Santo Thomas, Medan, Indonesia

⁴bang.heru82@gmail.com, ⁵saib@usu.ac.id, ⁶zakarias65@yahoo.com

Abstract. Affinity Propagation (AP) is exemplar-based clustering algorithm, this algorithm does not require prior knowledge of the number of clusters. The quality of clustering results is highly dependent on the “preference” value. Standard AP algorithm take “preference” value based on median or minimum value of similarity matrix, then the value is shared to all “preference” value on similarity matrix. This method does not give the best solution, because the value not represent the overall data structure. The Modified AP (M-AP) is proposed to resolve this problem. M-AP algorithm take “preference” value based on data distribution on each row from similarity matrix. Experimental result show that M-AP can outperform AP in quality clustering result based on Silhouette Index score.

1. Introduction

The process of grouping a set of physical or abstract object into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to another within the same cluster and dissimilar of data in other clusters [1]. Clustering techniques have been used in many fields, such as artificial intelligence, biology, data mining, machine learning, marketing, pattern recognition and others [2].

Affinity Propagation (AP) is a new clustering algorithm proposed by Brendan and Delbert Duek [3]. Unlike previous clustering algorithm such as k-means which taking random data points as first potential exemplars, AP considers all data points as potential cluster centers.

AP algorithm requires the value of “preference” parameter as the initial input, this “preference” parameter will directly affect the quality of clustering resulting by the AP algorithm [3]. AP algorithm take median (P_m) or minimum (P_{min}) value of the similarity matrix and shared that value as “preferences” value for all “preference” in the similarity matrix. The P_m will resulting in a moderate number of clusters and the P_{min} will resulting in a small number of clusters.

As [4] suggests that in many cases, setting a “preference” value based on the P_m value or P_{min} value for all “preference” values in the similarity matrix is not the best solution, because the P_m value or P_{min} value can't represent the overall data structure. Therefore, the determination of “preference” value becomes very important in AP algorithm, because the value will greatly affect the quality of AP algorithm itself [5].

2. Affinity Propagation

Affinity propagation (AP) is a new exemplar-based clustering algorithm proposed by Brendan and [3]. AP viewing all data points as a node in network, then the message exchanged recursively transmits along the edge of network until a good of exemplars emerges. Exemplar is the best data point to represent data clusters. Fig. 1 shown how the AP works.



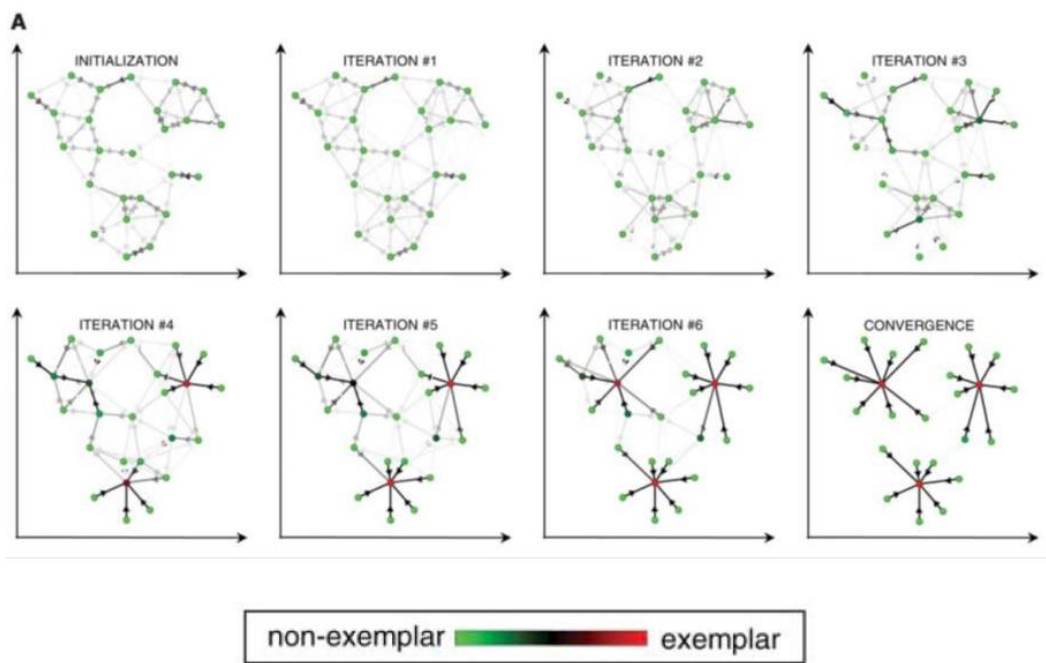


Figure 1. How affinity propagation works

AP take as input of real-valued similarities between data points, where $s(i, k)$ indicates how well the data point with index k is suited to be exemplar for data point i . Because the goal is to minimize squared error, each similarity in set to a negative square error (Euclidean distance), the similarity computed as:

$$s(i, k) \leftarrow -\|x_i - x_k\|^2 \quad (1)$$

2.1. Input preference

The “preference” value is the diagonal value of the similarity matrix, the default value of “preference” is computed from median (P_m) or the minimum (P_{min}) value of similarity matrix.

$$p = \text{median}(s(:)) \quad (2)$$

$$p = \min(s(:)) \quad (3)$$

This value then shared as “preference” value on similarity matrix, so every “preference” on similarity matrix has same value.

2.2. Messages passing

The process of AP can be viewed as a message passing process with two kinds of messages exchanged among data points: that message are responsibility and availability [6], these two kinds of messages can determine which points are served as exemplar and the point that belong of the exemplar [7]. Message passing process is shown in Fig. 2:

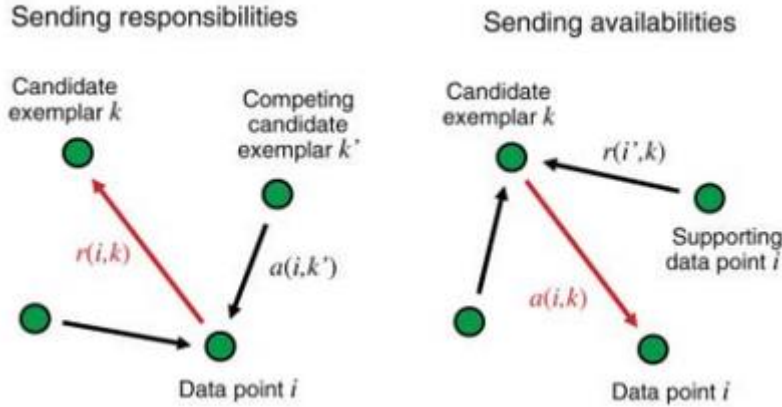


Figure 2. Message passing process

Responsibility, $r(i, k)$, is a message from data point i to k that reflects the accumulated evidence for how well-suited data point k is to serve as the exemplar for data point i . Responsibility, $r(i, k)$ computed as:

$$r(i, k) \leftarrow s(i, k) - \max_{k' \text{ s.t. } k' \neq k} \{a(i, k') + s(i, k')\} \quad (4)$$

Availability, $a(i, k)$, is a message from data point k to i that reflects the accumulated evidence for how appropriate it would be for data point i to choose data point k as its exemplar. Availability, $a(i, k)$ computed as (5) and “self-availability” $a(k, k)$ computed as (6).

$$a(i, k) \leftarrow \min\{0, r(k, k) + \sum_{i' \text{ s.t. } i' \neq i, k} \max\{0, r(i', k)\}\} \quad (5)$$

$$a(k, k) \leftarrow \sum_{i' \text{ s.t. } i' \neq k} \max\{0, r(i', k)\} \quad (6)$$

2.3. Exemplar decision

At any point during AP process, availabilities and responsibilities can be combined to identify exemplars. Data point i will serve as exemplar if $k = i$, otherwise i will belong as member of exemplar k . Exemplar decision computed as:

$$k \leftarrow \arg \max \{r(i, k) + a(i, k)\} \quad (7)$$

3. Modified AP (M-AP)

In this section will propose new algorithm call Modified AP (M-AP), this algorithm designed to solve AP limitation to determine the best “preference” value. in M-AP method, “preference” value is computed for each row in similarity matrix as:

$$p(i) \leftarrow - \left(\text{median}(s(i)) - \min(s(i)) \right) \quad (8)$$

4. Experimental Results

This section compares the clustering performance between AP algorithm and M-AP algorithm based on Silhouette Index (SI) [8] score. The SI score computed as:

$$s(i) \leftarrow \frac{(b(x_i) - a(x_i))}{\max(a(x_i), b(x_i))} \quad (9)$$

AP algorithm that using in tested are written and run in Python [9], and M-AP are modified version from the AP algorithm. To testing the performance both algorithm, the datasets used are from [10] as shown in table 1.

Table 1. Datasets

Dataset	Size	Class	Dimension
Wine	178	3	13
Iris	150	3	4
Yeast	1.484	10	8

The “preference” value for AP algorithm are computed using (2), whereas in the M-AP algorithm, the “preference” value is computed using (8). The result for both algorithm is shown in table 2.

Table 2. Experimental result

Dataset	Silhouettes Index Score	
	M-AP	AP
Wine	0.731	0.714
Iris	0.670	0.530
Yeast	0.271	0.263

5. Conclusions

This paper proposed new algorithm named M-AP, in this algorithm the “preference” value is computed each row from similarity matrix. Based on the experimental result that shown in table 2, M-AP algorithm can outperform AP algorithm, based on Silhouettes Index Score.

6. References

- [1] Han J and Kamber M 2006 Data Mining: Concepts and Techniques. 2nd Edition Elsevier: San Francisco.
- [2] Yadav J and Sharma M 2013 A Review of K-mean Algorithm International Journal of Engineering Trends and Technology (IJETT). 7: 2972-2976.
- [3] Frey BJ and Dueck D 2007 Clustering by Passing Messages Between Data Points Science. 315: 972-976.
- [4] Wang K, Zhang J, Li D, Zhang X and Guo T 2007 Adaptive Affinity Propagation Clustering. Acta Automatica Sinica. 33(12): 1242-1246.
- [5] Refianti R, Mutiara AB, Suhendra A and Juarna A 2017 Modified Adaptive Affinity Propagation with Similarity Distribution Based Preference. Proceedings of 2017 Second International Conference on Informatics and Computing (ICIC). pp. 1-5.
- [6] Fujiwara Y, Irie G and Kitahara T 2011 Fast Algorithm for Affinity Propagation Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence. pp. 2238-2243.
- [7] Ding L and Minghu J 2012 Affinity Propagation Clustering on Oral Conversation Texts Proceedings of 2012 IEEE 11th International Conference on Signal Processing. pp. 2279-2282.

- [8] Rousseeuw PJ 1986 Silhouettes: A Graphical Aid to The Interpretation and Validation of Cluster Analysis *Journal of Computational and Applied Mathematics*. 20(1987): 53-65.
- [9] <http://scikit-learn.org/stable/modules/generated/sklearn.cluster.AffinityPropagation.html>, 2018
- [10] <http://archive.ics.uci.edu/ml/>, 2018