A K-AP Clustering Algorithm Based on Manifold Similarity Measure

Hongjie Jia, Liangjun Wang, Heping Song, Qirong Mao, Shifei Ding

A *K*-AP Clustering Algorithm Based on Manifold Similarity Measure

Hongjie Jia[1], Liangjun Wang[1], Heping Song[1], Qirong Mao[1], Shifei Ding[2,3]

1 School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang 212013, China
jiahj@ujs.edu.cn
2 School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China
3 Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China

Abstract. *K*-AP clustering algorithm is a kind of affinity propagation (AP) clustering that can directly generate specified *K* clusters without adjusting the preference parameter. Similar to AP clustering algorithm, the clustering process of *K*-AP algorithm is also based on the similarity matrix. How to measure the similarities of data points is very important for *K*-AP algorithm. Since the original Euclidean distance is not suit for complex manifold data structure, we design a manifold similarity measurement and proposed a *K*-AP clustering algorithm based on the manifold similarity measure (MKAP). If two points lie on the same manifold, we assume that there is a path inside the manifold to connect the two points. The manifold similarity measure uses the length of the path as the manifold distance between the two points, so as to compress the distance of the data points in high-density region, while enlarge the distance of data points in low-density region. The clustering performance of the proposed MKAP algorithm is tested by comprehensive experiments. The clustering results show that MKAP algorithm can well deal with the datasets with complex manifold structures.

Keywords: *K*-AP clustering; similarity matrix; manifold similarity measure; affinity propagation

1 Introduction

Clustering is an important approach to analyze the intrinsic structure of data. Affinity propagation (AP) clustering, proposed by Frey and Dueck [1], is a popular clustering method. AP clustering aims to find the optimal representative point, called 'exemplar', for each data point. It is more useful to find representative points than separate date points into several classes in many application domains [2-5]. For example, the representative points recognized from a document can be used to summarize and refine an essay. Different from k-means, the AP algorithm does not need specifying the initial

adfa, p. 1, 2011.

cluster centers in advance [6,7]. In contrast, it regards all data points as potential cluster center, therefore avoiding the arbitrary of the selection of the initial cluster centers.

However, AP clustering algorithm cannot directly specify the final class number, and the number of ultimate clusters is affected by a user-defined parameter. In order to generate K clusters, Zhang, et al. [8] propose K-AP clustering algorithm. Similar to AP algorithm, K-AP algorithm needs constructing similarity matrix firstly, so it is crucial to select an appropriate distance measurement to describe the real structure of dataset. The data points belong to the same cluster should have high similarity, and keep the spatial coherency [9]. K-AP algorithm has better clustering performance on linear separable data, but not suit the clustering problem of manifold data. Because K-AP algorithm measures the similarity between data points based on Euclidean distance, which cannot correctly reflect the distribution of complex manifold data set [10]. This will significantly reduce the performance of K-AP, causing bad clustering results. According to the assumption of local-coherence and global-coherence of cluster, this paper designs a manifold similarity measure. We use a density-adjustable length to calculate the distance of data points, so that it is able to describe the manifold data distribution much better. Then the manifold similarity measure is used to improve the performance of K-AP algorithm.

To solve the difficulties of handling manifold data faced by K-AP clustering algorithm, we propose a K-AP clustering algorithm based on manifold similarity measure (MKAP). The rest paper is organized as follows: Section 2 introduces the basic theory of K-AP Clustering algorithm; Section 3 describes the manifold similarity measure; Section 4 presents the MKAP algorithm and gives its detail steps; Section 5 verifies the effectiveness of MKAP algorithm on artificial data sets and real world data sets; the last part is conclusion.

2 Basic *K*-AP Clustering

In AP clustering algorithm, the cluster number is affected by the preference parameter. It is not easy to set an appropriate preference parameter for AP algorithm to get the desired number of clusters [11]. *K*-AP clustering algorithm solves this problem very well. It uses the specified cluster number *k* as an input parameter and can directly classifies data points into *k* groups. *K*-AP algorithm searches the optimal representative point set of clusters and maximize the energy function by passing messages between data points. Equation (1) is the energy function of *K*-AP algorithm:

$$E(\varepsilon) = \sum_{j=1}^{K} \sum_{x_i : c(x_i) = \mathrm{e}_j} s(x_i, e_j) \tag{1}$$

where K is the cluster number and the number of representative points; $\varepsilon = \{e_1, \cdots, e_k\}$ is the collection of representative points; $c(x_i)$ is the mapping function between x_i and its closest representative point; $s(x_i, e_j)$ is the similarity between x_i and cluster representative point e_j.

To find K representative points, we may introduce binary variables $\left\{b_{ij} \in \{0,1\}, i, j = 1, \cdots, N\right\}$ to indicate the distribution of representative points: $b_{ij} = 1$, $i \neq j$ means x_i chooses x_j as its representative point; $b_{ii} = 1$ means x_i is a representative point. Then Equation (1) is equal to Equation (2):

$$E(\{b_{ij}\}) = \sum_{i=1}^{N}\sum_{j=1}^{N} b_{ij} s(x_i, x_j) \tag{2}$$

Equation (2) satisfies three conditions: $\sum_{j=1}^{N} b_{ij} = 1$; $b_{ii} = 1$, if $\exists b_{ji} = 1$; $\sum_{i=1}^{N} b_{ii} = K$. The three conditions mean that: a) every x_i can only have one representative point; b) if there is a point x_j select x_i as its representative point, then x_i is a representative point; c) the number of representative points must be K. These constraint conditions can be solved by factor graph model. Then the problem of finding K representative points turns into searching the optimal value of b_{ij} in factor graph. Equation (3) is the objective function of K-AP:

$$\begin{aligned} &F(b; s; K) \\ &= \prod_{i=1}^{N}\left(e^{b_{ii}} \prod_{j=1, j\neq i}^{N} e^{b_{ij}s(i,j)}\right) h(b_{11}, \cdots, b_{NN} \mid K) \prod_{j=1}^{N} f_j(b_{1j}, \cdots, b_{Nj}) \prod_{i=1}^{N} g_i(b_{i1}, \cdots, b_{iN}) \end{aligned} \tag{3}$$

where $\{g_i\}$, $\{f_i\}$ and h are three constraint functions. The above linear programming problem can be solved by Belief Propagation (BP) method [8].

3 Manifold Similarity Measure

The standard K-AP clustering algorithm measures the similarity between data points by Gaussian kernel function. Gaussian kernel is based on Euclidean distance, but Euclidean distance is not a proper distance measure for manifold data. Figure 1 is an example to illustrate the shortcomings of Euclidean distance.

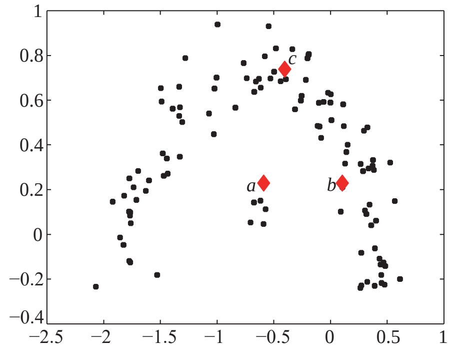


Figure 1. Euclidean distance for manifold data

It can be seen from Figure 1 that point b and point c are on the same manifold, point a and point b are on different manifolds. We hope that the similarity between point b and point c is greater than the similarity between point a and point b, so that it is possible to group b and c into the same cluster. However, the Euclidean distance between point a and point b is significantly smaller than the Euclidean distance between point b and point c. We assume that the similarity of data pairs in the same manifold structure is high, and the similarity of data pairs in different manifold structures is low [12]. So this paper presents a manifold similarity function to meet the clustering assumption. First we define a segment length in manifold data.

Definition 1. The length of line segment on manifold:

$$L(x, y) = e^{\rho d(x,y)} - 1 \tag{4}$$

where $d(x, y) = \|x - y\|$ is the Euclidean distance between the data points x and y; ρ is called the scaling factor.

If two points lie on the same manifold, suppose there is a path inside the manifold to connect the two points. We can use the length of the path as the manifold distance between the two points [13]. According to the length of line segment on manifold, a new distance measure—manifold distance measure is defined in Definition 2.

Definition 2. Manifold distance measure: Given an undirected weighted graph $G = (V, E)$, let $p = \{v_1, v_2, \ldots, v_{|p|}\} \in V^{|p|}$ denote the path between vertex v_1 and $v_{|p|}$, where $|p|$ is the number of vertices contained in path p, the edge $(v_k, v_{k+1}) \in E$, $1 \leq k < |p|$. Let P_{ij} represent the set of all paths connecting the point pair $\{x_i, x_j\}$ $(1 \leq i, \ j < N)$, then the manifold distance between x_i and x_j is

$$D_{i,j}^{\rho} = \frac{1}{\rho^2} \ln(1 + d_{sp}(x_i, x_j))^2 \tag{5}$$

where $d_{sp}(x_i, x_j) = \min_{p \subset P_{ij}} \sum_{k=1}^{|p|-1} L(v_k, v_{k+1})$ is the distance of the shortest path between nodes x_i and x_j on graph G; $L(v_k, v_{k+1})$ is the manifold segment distance of two adjacent points on the shortest path from x_i to x_j on graph G.

Definition 3. According to the above manifold distance measure, the manifold similarity of data points x_i and x_j is defined as

$$s(i, j) = \exp\left(-\frac{D_{i,j}^{\rho}}{2\sigma_i \sigma_j}\right) \tag{6}$$

where the scale parameter $\sigma_i = d(x_i, x_{il}) = \|x_i - x_{il}\|$, x_{il} is the l-th neighbor of x_i. σ_i adaptively changes with the neighborhood distribution of data points. The manifold similarity can enlarge the distance between two points on different manifolds and reduce the distance between two points on the same manifold.

4 *K*-AP Clustering Based on Manifold Similarity Measure

We use the manifold similarity measure to improve the *K*-AP clustering algorithm, and proposes a MKAP algorithm. This algorithm constructs the similarity matrix with the manifold similarity measure. Then it iteratively optimizes the clustering objective function by passing messages. The detail steps of MKAP algorithm are given below.

Algorithm 1. *K*-AP clustering algorithm based on manifold similarity measure

Input: data set $X = \{x_1, x_2, \ldots, x_n\}$, cluster number k.

Output: k final clusters.

Step 1. Calculate the manifold distance $D_{i,j}^{\rho}$ between each data pair (x_i, x_j) according to Equation (5).

Step 2. Use the manifold distance $D_{i,j}^{\rho}$ to calculate the similarity $s(i, j)$ between pairwise points (x_i, x_j) by Equation (6), and construct the similarity matrix S.

Step 3. Initialize the 'availability' $a(i, j) = 0$, and the 'confidence' $\eta^{out}(i) = \min(S)$.

Step 4. Iteratively update the 'responsibility', 'availability' and 'confidence' according to the following equations:

1) Update the 'responsibility', $\forall i, j$:

$$r(i, j) = s(i, j) - \max\left\{\eta^{out}(i) + a(i,i), \max_{j':j'\notin\{i,j\}}\left\{s(i, j') + a(i, j')\right\}\right\} \tag{7}$$

$$r(i,i) = \eta^{out}(i) - \max_{j':j'\neq s}\left\{s(i, j') + a(i, j')\right\} \tag{8}$$

2) Update the 'availability', $\forall i, j$:

$$a(i, j) = \min\left\{0, r(j, j) + \sum_{v':v'\notin\{i,j\}} \max\left\{0, r(i', j)\right\}\right\} \tag{9}$$

$$a(j, j) = \sum_{v':v'\neq j} \max\left\{0, r(i', j)\right\} \tag{10}$$

3) Update the 'confidence', $\forall i$:

$$\eta^{in}(i) = a(i,i) - \max_{j':j'\neq s}\left\{s(i, j') + a(i, j')\right\} \tag{11}$$

$$\eta^{out}(i) = -f^{k}\left(\left\{\eta^{in}(j), j \neq i\right\}\right) \tag{12}$$

where $f^k(\bullet)$ means the k-th largest value in $\eta^{in}(j)$, $i, j = 1, 2, \ldots, N$.

Step 5. According to Equation (13) to determine the best cluster center for data points, until the algorithm converges.

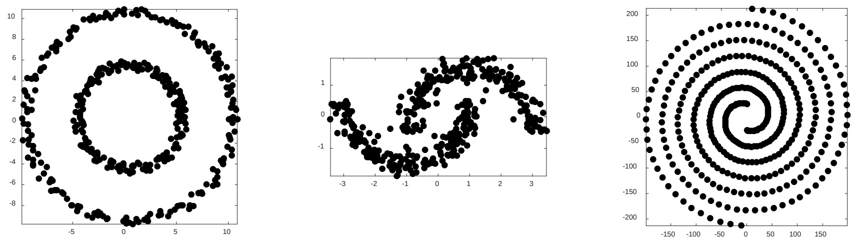
$$c_i = \arg\max_j\left\{a(i, j) + r(i, j)\right\} \tag{13}$$

Similar to K-AP algorithm, the time complexity of MKAP algorithm is also $O(N^2)$. As MKAP algorithm uses the manifold similarity measure to construct the similarity matrix, it can well describe the manifold relationship between data points.

5 Experimental Analysis

5.1 Clustering on Synthetic Datasets

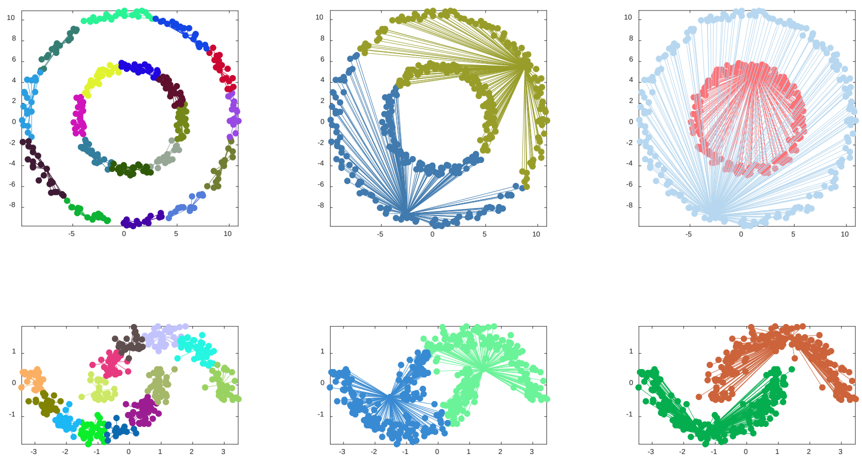
In the experiments, the clustering performances of AP algorithm, K-AP algorithm and MKAP algorithm are compared on three challenging synthetic manifold datasets: 'two circles', 'two moons' and 'two spirals'. These datasets are illustrated in Figure 2.

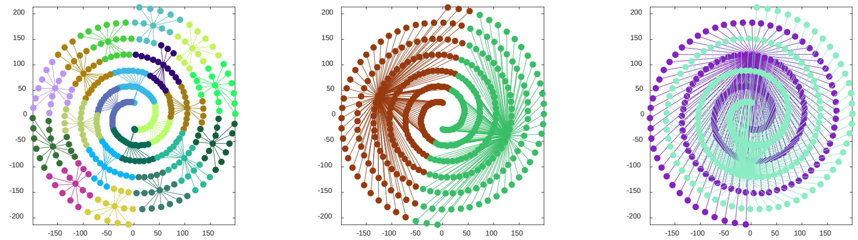


(a) two circles (b) two moons (c) two spirals

Figure 2. Original synthetic datasets

In the experiments, the preference parameter p of AP algorithm is the median of affinity matrix, the maximum iteration maxits = 1000, the convergence coefficient of iteration convits = 100. The density factor of MKAP algorithm is set as ρ = 2. The clustering results of AP algorithm, K-AP algorithm and MKAP algorithm on these three synthetic data sets are presented in Figure 3.





(a) AP algorithm (b) *K*-AP algorithm (c) MKAP algorithm

Figure 3. Clustering results of different algorithms on synthetic datasets

Form Figure 3, we can see that AP algorithm tends to generate many small clusters. It is not easy to control the cluster number of clustering results for AP algorithm. AP algorithm is easy to fall into the local optimum. In *K*-AP algorithm, the cluster number *K* is one of the clustering constraints, so the final cluster number of *K*-AP algorithm on each dataset is right. But similar to AP algorithm, *K*-AP algorithm measures the similarity between points based on Euclidean distance and it cannot recognize complex manifold structure of the dataset. In contrast, the performance of the proposed MKAP algorithm is much better. With the help of manifold similarity measurement, MKAP algorithm is suitable for the clustering problem on manifold datasets. For MKAP algorithm, the data points on the same manifold have high similarity and the data points on different manifolds are dissimilar with each other.

5.2 Clustering on Real World Datasets

To further test the effectiveness of the proposed MKAP algorithm, we compare MKAP algorithm with other popular clustering algorithms on several benchmarking real world datasets [14]. The information of these datasets are shown in Table 1.

Table 1. Information of real world datasets

Data set	Number of objects	Number of attributes	Number of classes
Dermatology	336	34	6
Ionosphere	351	34	2
Sonar	208	60	2
WDBC	569	30	2
Wine	178	13	3
Zoo	101	16	7

In the experiments, adjusted rand index (ARI) is used to evaluate the clustering performance [15]. ARI is based on the relationship of pairwise data points. The calculation equation of ARI is:

$$\mathrm{ARI} = \frac{2(a*d-b*c)}{(a+b)*(b+d)+(a+c)*(c+d)} \tag{14}$$

where a, b, c, d are the number of different kind of data pairs. ARI $\in [0,1]$, the higher the value of ARI, the better the clustering quality.

The clustering performance of the proposed MKAP algorithm is compared with AP algorithm, K-AP algorithm and F-AP algorithm [16]. All the experiments are conducted on the computer with 3.20 GHz AMD Ryzen 5 1600 six-core processor, 8GB RAM. The programming environment is MATLAB 2015b. The clustering results of different algorithms are given in Table 2.

Table 2. Clustering results of different algorithms on real world datasets

Dataset	Evaluation index	Algorithm			
		AP	*K*-AP	F-AP	MKAP
Dermatology	ARI index	0.1427	0.0405	0.0331	**0.1718**
	Time (s)	2.1755	3.6454	0.1611	4.1073
	Cluster number	16	6	10	6
Ionosphere	ARI index	0.1208	0.1728	0.1776	**0.1867**
	Time (s)	1.8923	1.4956	0.1525	1.6541
	Cluster number	41	2	3	2
Sonar	ARI index	0.0206	0.0011	0.0064	**0.0287**
	Time (s)	1.9220	1.8528	0.6148	2.1602
	Cluster number	23	2	7	2
WDBC	ARI index	0.0963	0.2787	0.0677	**0.3214**
	Time (s)	3.5996	4.5231	1.3536	3.3812
	Cluster number	21	2	16	2
Wine	ARI index	0.2073	0.3465	**0.3711**	0.3316
	Time (s)	1.4214	0.8184	0.5734	1.0361
	Cluster number	8	3	3	3
Zoo	ARI index	0.5158	0.6486	0.6690	**0.7324**
	Time(s)	2.8031	2.2104	0.5629	2.6133
	Cluster number	8	7	8	7

According to Table 2, the running speed of F-AP algorithm is much faster than other algorithms. Because F-AP computes upper and lower estimates to limit the messages to be updated in each iteration, and it dynamically detects converged messages to efficiently skip unneeded updates. But it is not easy for AP algorithm and F-AP algorithm to control the final cluster number. Their clustering performance are not very well on some datasets. Both K-AP algorithm and MKAP algorithm can make good use of prior knowledge, and divide dataset into a given number of clusters. However, K-AP constructs the similarity matrix based on the Euclidean distance between data points. Euclidean distance is not proper to describe the complex data structure of many real world datasets. So the ARI indexes of K-AP algorithm are not as good as the proposed MKAP algorithm on most datasets. MKAP utilizes the manifold similarity measure to do clustering and can produce better clustering results.

6 Conclusions

In this paper, we propose a *K*-AP clustering algorithm based on manifold similarity measure (MKAP). *K*-AP algorithm cannot work well on manifold data and it is easy to fall into local optimum. To improve the clustering performance *K*-AP algorithm, we design a manifold similarity measurement. The manifold similarity measure can correctly describe the complex relationships between data points and reveal the internal structure of the dataset. With the manifold similarity measure, MKAP algorithm is able to maintain the global and local consistency of clustering when assigning data points into multiple groups. In the experiments, the proposed MKAP algorithm is compared with other popular Affinity propagation clustering algorithms on both synthetic and real world datasets. The experimental results demonstrate the effectiveness of MKAP algorithm. Next we consider to improve the clustering efficiency of MKAP algorithm and apply it to some practical problems, such as character recognition, image segmentation and speech separation etc.

Acknowledgements

This work is supported by the National Natural Science Foundations of China (Nos. 61672267, 61672522, 61601202), and the Natural Science Foundation of Jiangsu Province (Nos. BK20140571, BK20170558).

References

1. Frey BJ, Dueck D. Clustering by Passing Messages Between Data Points. Science, 2007, 315(5814): 972-976.
2. Wei Z, Wang Y, He S, et al. A novel intelligent method for bearing fault diagnosis based on affinity propagation clustering and adaptive feature selection. Knowledge-Based Systems, 2017, 116: 1-12.
3. Jia H, Ding S, Du M. A Nyström spectral clustering algorithm based on probability incremental sampling. Soft Computing, 2017, 21(19): 5815-5827.
4. Wang Z J, Zhan Z H, Lin Y, et al. Dual-Strategy Differential Evolution with Affinity Propagation Clustering for Multimodal Optimization Problems. IEEE Transactions on Evolutionary Computation, 2017. (DOI: 10.1109/TEVC.2017.2769108)
5. Li P, Gu W, Wang L, et al. Dynamic equivalent modeling of two-staged photovoltaic power station clusters based on dynamic affinity propagation clustering algorithm. International Journal of Electrical Power & Energy Systems, 2018, 95: 463-475.
6. Li P, Ji H, Wang B, et al. Adjustable preference affinity propagation clustering. Pattern Recognition Letters, 2017, 85: 72-78.
7. Fan Z, Jiang J, Weng S, et al. Adaptive density distribution inspired affinity propagation clustering. Neural Computing and Applications, 2017: 1-11. (DOI: 10.1007/s00521-017-3024-6)
8. Zhang XL, Wang W, Nørvag K, et al. K-AP: Generating Specified K Clusters by Efficient Affinity Propagation. Proceedings 2010 10th IEEE International Conference on Data Mining (ICDM 2010), 2010: 1187-1192.

9. Jia H, Ding S, Du M. Self-tuning p-spectral clustering based on shared nearest neighbors. Cognitive Computation, 2015, 7(5): 622-632.
10. Wang B, Zhang J, Liu Y, et al. Density peaks clustering based integrate framework for multi-document summarization. CAAI Transactions on Intelligence Technology, 2017, 2(1): 26-30.
11. Arzeno NM, Vikalo H. Semi-supervised affinity propagation with soft instance-level constraints. IEEE transactions on pattern analysis and machine intelligence, 2015, 37(5): 1041-1052.
12. Liu Z, Wang W, Jin Q. Manifold alignment using discrete surface Ricci flow. CAAI Transactions on Intelligence Technology, 2016, 1(3): 285-292.
13. Jia H, Ding S, Xu X, et al. The latest research progress on spectral clustering. Neural Computing & Applications, 2014, 24(7-8): 1477-1486.
14. Jia H, Ding S, Du M, et al. Approximate normalized cuts without Eigen-decomposition. Information Sciences, 2016, 374: 135-150.
15. Jia H, Ding S, Meng L, et al. A density-adaptive affinity propagation clustering algorithm based on spectral dimension reduction. Neural Computing & Applications, 2014, 25(7-8): 1557-1567.
16. Fujiwara Y, Nakatsuji M, Shiokawa H, et al. Adaptive message update for fast affinity propagation. Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 2015: 309-318.