Affinity Propagation Clustering With Pairwise Constraints

Lijia Zhang[a], Lianglun Cheng

School of Computer Science and Technology，Guangdong University of Technology，Guangzhou 510006，China.

[a]781176915@qq.com

Keywords: Affinity propagation, complex dataset, semi-supervised method, pairwise constraints.

Abstract. With the explosive growing of data, there are challenges to deal with the large scale complex data. Many clustering algorithms have been proposed. Such as Affinity Propagation (AP) clustering Algorithm, AP takes similarity between pairs of data point as input measures. AP is a fast and efficient clustering algorithm for large dataset compared with the existing clustering algorithm. But for the datasets with complicated cluster structure, it cannot produce good clustering results. It can improve the clustering effect of AP by using the pairwise constraints and extended pairwise constraints to adjust the similarity matrix. Therefore, a semi-supervised method of affinity propagation clustering with pairwise constraints (AP with PC) is proposed in this paper. Experiments show that the method has good clustering result for complex datasets, moreover, the method is better than the comparative algorithm when the number of constraints for is large.

1. Introduction

Clustering algorithm has been widely applied to data mining, pattern classification and many other fields. With the rapid development of information and network technology, data size is increasing and data type is being complex, and the requirements of clustering efficiency and effect are also getting higher. For improving the efficiency and effectiveness of large-scale complex data clustering, AP clustering algorithm[1] has been proposed.

The method of selecting the representatives of traditional clustering algorithm is that representatives are randomly selected firstly, and then representatives are being adjusted by calculating iteratively until representatives are no longer obviously changed or the iteration is completed. The selection of the initial representatives is connected with clustering results. AP clustering algorithm take all data points of a dataset as possible representatives and then calculate Responsibility and Availability iteratively based on the similarity matrix to adjust representatives. Moreover, AP clustering algorithm is a fast and efficient clustering algorithm for large dataset.

In many practical problems, because there is not any priori analysis of data, traditional clustering algorithms can not get effective clustering results sometimes. However, in some practical problems, we can get less of priori knowledge of datasets, including class labels and constraint conditions (such as pairwise constraints). Thus, how to use a small amount of prior knowledge for clustering the large amount of data without prior knowledge has become a very important problem.

In this paper, the original AP clustering algorithm are combined with semi-supervised method[2-4], the similarity matrix is adjusted by introducing the pairwise constraints heuristically. The clustering effect is improved by clustering on the new adjusted similarity matrix. Experimental results show that the effect of AP with PC has significantly improved compared with the original clustering algorithm.

2. Affinity Propagation Clustering Algorithm

Frey and Dueck first proposed AP clustering algorithm in Science. AP clustering algorithm takes similarity between pairs of data point as input measures. AP clustering algorithm is a fast and efficient clustering algorithm for large dataset compared with the existing clustering algorithm. AP clustering algorithm take all data points of a dataset as possible representatives and then calculate Responsibility and Availability iteratively based on the similarity matrix to adjust representatives. AP

clustering algorithm has 3 indicators, namely Similarity, Responsibility and Availability, defined as follows:

Similarity: Assuming that $X=\{x_1,x_2,...,x_n\}$ is a dataset with n samples, the similarity of the dataset is:

$$s(i,j)=-\left\|x_i-x_j\right\|^2, i\neq j \tag{1}$$

In equation 1, $s(i,j)=0$ represents that x_i and x_j has a greatest similarity, and $s(i,j)=-\infty$ represents that x_i and x_j belong to different categories. $s(i,i)$ represents preference parameter, $s(i,i)$ is greater, the probability that x_i is selected as an exemplar is higher. Preference parameter also indicates the number of exemplars, and the larger is the parameter, the more is exemplars. Another two indicators of the AP algorithm are:

Responsibility: $r(i,k)$ sent from x_i to candidate exemplar x_k, reflects the accumulated evidence for how well-suited x_k is to serve as the exemplar for x_i, taking into account other potential exemplars for x_i.

Availability: $a(i,k)$ send from candidate exemplar x_k to x_i, reflects the accumulated evidence for how appropriate it would be for x_i to choose point k as its exemplar, taking into account the support from other points that x_k should be an exemplar.

AP algorithm calculates the above two indicators iteratively based on the similarity matrix, as follows:

$$r(i,k)=\begin{cases} s(i,k)-\max\limits_{k'\neq k}\left[a(i,k')+s(i,k')\right], i\neq k \\ s(k,k)-\max\limits_{k'\neq k}\left[s(k',k)\right], i=k \end{cases} \tag{2}$$

$$a(i,k)=\begin{cases} \min\limits_{i\neq k}\left\{0,r(k,k)+\sum\limits_{i'\neq i,i'\neq k}\max\left[0,r(i',k)\right]\right\}, i\neq k \\ \sum\limits_{k'\neq k}\max\left[0,r(k',k)\right], i=k \end{cases} \tag{3}$$

In order to avoid the occurrence of shock, damping parameter $\lambda\in[0,1)$ is introduced. The updating of Responsibility and Availability are as follows:

$$r_{new}(i,k)=\lambda r_{old}(i,k)+(1-\lambda)r(i,k) \tag{4}$$

$$a_{new}(i,k)=\lambda a_{old}(i,k)+(1-\lambda)a(i,k) \tag{5}$$

In equation 4 and 5, the bigger is λ, the better is the effect of the elimination of shock, but the convergence rate of the algorithm will be slow.

3. Affinity Propagation Clustering With Pairwise Constraints

AP algorithm can not meet the requirements of complex data clustering, in order to solve the problem, the semi-supervised method of AP clustering with pairwise constraints is proposed in this paper. In the AP algorithm, the similarity matrix and the preference parameter are two important parameters. The preference parameter is independent in a datasets, thus it is difficult to determine the independent parameters by using a priori information. Therefore, it is reasonable to use pairwise constraints to adjust the similarity matrix. Semi-supervised clustering has two pairwise constraints, which are Must-link and Cannot-link. Must-link constraint requires two data points must be the same class, while the Cannot-link constraint requires two data points to be not the same class. Assuming that M and C are respectively representing the Must-link constraint set and the Cannot-link constraint set, the similarity matrix is adjusted as follows:

When x_i and x_j are in the same class, will meet the requirements of Must-link,

$$(x_i,x_j)\in M\Rightarrow s(i,j)=0\,\&\,s(i,j)=0 \tag{6}$$

When x_i and x_j are not in the same class, will meet the requirements of Cannot-link,

$$(x_i, x_j) \in C \Rightarrow s(i,j) = -\infty \,\&\, s(i,j) = -\infty \quad (7)$$

In addition to the known constraints that adjust the similarity matrix, similarity of other data points also need to adjust. Given the following two extensions of constraints:

If a data point and many other data points satisfy Must-link constraints at the same time, then these data points are Must-link constraints each other,

$$(x_i, x_j) \in M \,\&\, (x_j, x_k) \in M \Rightarrow (x_i, x_k) \in M \quad (8)$$

If x_i and x_j satisfy Cannot-link constraints, and then x_i and all data points that meet the requirements of Must-link constraints with x_i satisfy Cannot-link constraints,

$$(x_i, x_j) \in C \,\&\, (x_j, x_k) \in M \Rightarrow (x_i, x_k) \in C \quad (9)$$

After the above adjustments, the similarity matrix has changed largely. AP algorithm is based on the similarity matrix, so the whole iterative process can be changed. AP clustering with pairwise constraints can spread the pairwise constraint information through the nearest neighbor information, while affect the clustering results.

4. Experimental Results and Analysis

The environment of this experiment is: CPU Intel dual core 2.5GHz, memory 4GB, hard disk for the 128GB SSD, operating system is 64 bit Windows 7.

The datasets used in this experiment is the datasets of UCI database[5], including Iris dataset, Wine dataset and Glass dataset. The datasets are shown in Table 1.

Table 1 List of Datasets

Datasets	Number of data points	Number of Attributes	Number of cluster centers
Iris	150	4	3
Wine	178	13	3
Glass	214	9	6

The evaluation indicators used in this study are CRI and F_{CRI} which are defined as follows,

CRI is used to evaluate the effect of semi-supervised clustering algorithm, and it is defined as the data points from the same cluster:

$$CRI = Cfd / Tfd \quad (10)$$

In equation 10, $Tfd = (n(n-1)/2 - C_n)$, n is the number of data points, C_n is the number of pairwise constraints, Cfd is the number of the correct number of the partition of the correct data for the division of the number of pairs of constraints.

F_{CRI} is the ratio of the correct implementation of the Must-linked constraint and Cannot-linked constraint information. And F_{CRI} is evaluated according to the correct clustering results of the same kind of data points and the definitions are follows:

$$R_{ML} = Cml / Tml \quad (11)$$

$$R_{CL} = Ccl / Tcl \quad (12)$$

$$F_{CRI} = 2 * R_{ML} * R_{CL} / (R_{ML} + R_{CL}) \quad (13)$$

Cml is the number of Must-linked constraints that are implemented correctly, *Tml* is the total number of Must-linked constraints, *Ccl* is the number of Cannot-linked constraints that are implemented correctly, *Tcl* represents the total number of Cannot-linked constraints. F_{CRI} is the balance of R_{ML} and R_{CL}.

To the accuracy of the experiment, this paper test each dataset 20 times under a given number of constraint information, and then take average value represents the algorithm under a given number of constraint information for a dataset clustering effect. Meanwhile, in AP algorithm, the preference parameters set as the average value of the similarity, the damping parameter is 0.5, and the iteration number is 400. In this experiment, AP clustering with pairwise constraints (AP with PC) is contrasted

with K-means clustering with pairwise constraints[6] (K-means with PC) and AP. Experimental results are shown in Figure 1 and Figure 2.

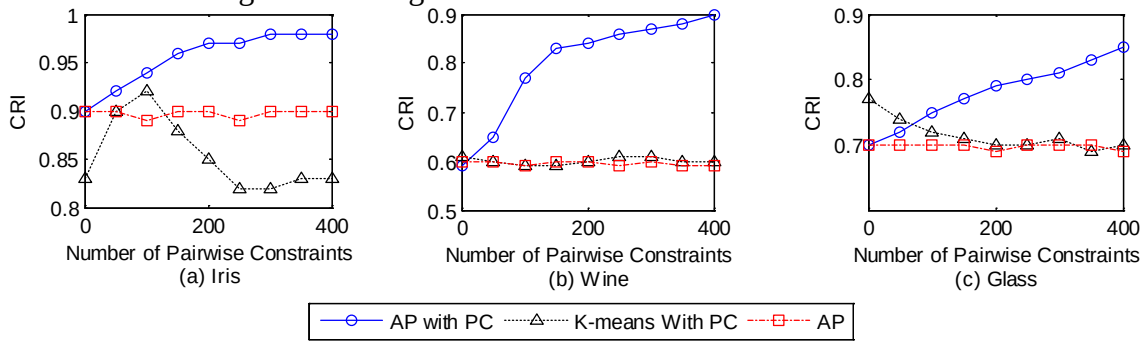

Fig. 1 CRI evaluation results of the algorithms

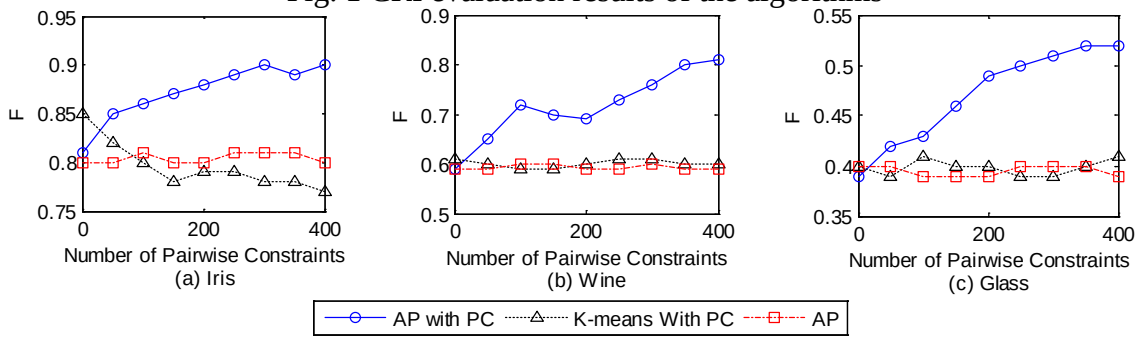

Fig. 2 F_{CRI} evaluation results of the algorithms

In Fig. 1 and Fig. 2, AP with PC gives the better clustering results. For the both two evaluation index, with the increase of constraints, the effect of the algorithm is uptrend, and on the sufficient constraints AP with pairwise constraints is better than other clustering algorithms. Moreover, AP with PC is more stability and effect than K-means with PC, because the structure of similarity matrix is not variable in AP with PC. On the whole, the effect of AP with PC is better than other alignment algorithms.

5. Summary

In this paper, pairwise constraints and extended pairwise constraints are used in AP clustering algorithm, the similarity matrix is adjusted by introducing the pairwise constraints heuristically. The clustering effect is improved by clustering on the new adjusted similarity matrix. Experimental results show that the effect of AP with PC has significantly improved compared with the original clustering algorithm.

References

[1]. Frey B J, Delbert D. Clustering by passing messages between data points.[J]. Science, 2007, 315(5814):972-6.

[2]. Yu X. Semi-Supervised Clustering Based on Affinity Propagation Algorithm[J]. Journal of Software, 2008, 19(11):2803-2813.

[3]. Arzeno N M, Haris V. Semi-Supervised Affinity Propagation with Soft Instance-Level Constraints.[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 37(5):1041-1052.

[4]. Meng F. A Fast Semi-supervised Affinity Propagation Community Detection Algorithm[J]. Journal of Information & Computational Science, 2015, 12(8):3261-3274.

[5]. Information on: http://archive.ics.uci.edu/ml/datasets.html

[6]. Wang X, Wang C, Shen J. Semi–supervised K-Means Clustering by Optimizing Initial Cluster Centers[J]. Lecture Notes in Computer Science, 2011, 6988:178-187.