Pattern Recognition Letters 85 (2017) 72–78

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Adjustable preference affinity propagation clustering

Ping Li, Haifeng Ji*, Baoliang Wang, Zhiyao Huang, Haiqing Li

State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, PR China

ARTICLE INFO

Article history:
Received 15 May 2016
Available online 29 November 2016

Keywords:
Pattern recognition
Clustering analysis
Affinity propagation
Factor graph

ABSTRACT

A new Affinity Propagation (AP) algorithm, Adjustable Preference Affinity Propagation (APAP) algorithm, is proposed in this work. The distinguishing features of APAP algorithm are that the initial value of each element preference p_k is independently determined according to the data distribution and p_k will be automatically adjusted during the iteration process. Experiments on synthetic data and real data are carried out. Experimental results verified the effectiveness of the proposed APAP algorithm. Compared with the standard AP algorithm, APAP algorithm has a better overall performance and can obtain better clustering results.

1. Introduction

Affinity Propagation (AP) is a new exemplar-based clustering method proposed by Frey and Dueck in 2007 [1]. Compared with other methods, its distinguishing feature is that AP considers all the data points as potential exemplars and identifies clusters automatically (only the inputting similarities between data points are necessary) [1–5]. AP can avoid many poor clustering solutions caused by unlucky initializations and hard decisions [2,6]. Besides, the introduction of message-passing techniques in factor graph makes its computation efficiency higher, especially for large data sets [7–11]. Experiments in references [1,2,6,12–15] have shown that AP or AP-based methods can obtain better solutions (lower error and less time) than the previous algorithms (such as the mixtures of Gaussians [4], the K-Means algorithm [6,12,15], the K-Medoids algorithm [2,13], the spectral clustering [13,14] and the hierarchical clustering [2,13], etc.). Because of the above advantages, AP algorithm has become an attractive clustering method and has been studied/applied in many domains, including face recognition, image segmentation and categorization, gene expression signatures' extraction, text mining, facility location and so on [2,6,10–13,15–17].

Many improvement methods and extensions have been developed [10,11,18–23]. In 2007 and 2008, Leone et al. proposed a kind of Soft-Constraint AP (SCAP) algorithm, which relaxed the hard constraints of the standard AP algorithm, and used the SCAP algorithm to cluster the irregularity shaped data sets [10,18,19]. In 2008, Zhang extended AP to data steaming, which is significant in handling large-scale data sets [26]. In 2009, Givoni and Frey took account for instance-level constraints in the clustering process and proposed a Semi-Supervised AP algorithm (SSAP) [20]. In 2011, Givoni et al. proposed a Hierarchical AP (HAP) algorithm to solve hierarchical clustering problems [22]. In 2013, Wang et al. extended the single-exemplar model and developed a new Multi-Exemplar AP (MEAP) algorithm to cluster multi-subclasses data sets [23]. In 2014, Sun et al. proposed two new kinds of Incremental AP (IAP) algorithms (IAP clustering based on K-Medoids (IAPKM) and IAP clustering based on Nearest Neighbor Assignment (IAPNA)) to solve incremental clustering problems [11]. In 2014, Chen et al. proposed a kind of Stability-based AP (SAP) in which the preference was determined by the clustering stability [25]. In 2015, Arzeno and Vikalo assigned a confidence to the set of instance-level constraints and proposed Soft-Constraint Semi-Supervised AP (SCSSAP) [21]. In 2016, Wang et al. extended the single-view affinity propagation into multi-view affinity propagation (MVAP) algorithm, which implements the clustering by passing messages both within individual views and across different views [27].

Although many achievements have been obtained, AP is still a developing clustering method. The conventional AP or AP-based algorithms still have a drawback, i.e. how to determine the values of preference P (a vector $\{p_k\}_{N\times 1}$ whose elements indicate how likely the relevant data point is to be chosen as an exemplar) [2,3,8–10,24,25]. In most of AP or AP-based algorithms, all the values of element preferences are commonly set as a constant. It has been found that the preference significantly affects AP clustering result in many experiments [14,18,29]. Lower values may lead to fewer clusters, while the effect with higher values may be opposite [2,21,28]. Improper values may also lead to suboptimal clustering solutions [14,29]. Meanwhile, for practical clustering problems,

* Corresponding author. Fax: +86 571 87951219.
E-mail address: hfji@iipc.zju.edu.cn (H. Ji).

http://dx.doi.org/10.1016/j.patrec.2016.11.017
0167-8655/

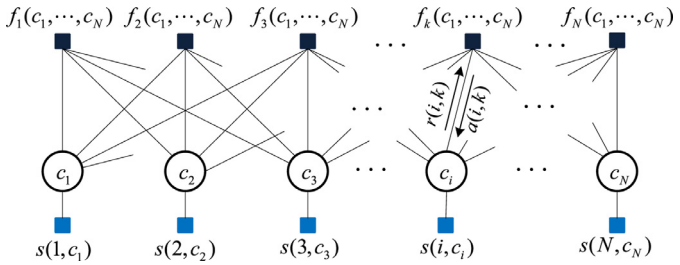

Fig. 1. A typical factor graph model of AP.

not every point has the same probability to become an exemplar. Making all the data points share one common exemplar preference may not be appropriate. It ignores the information behind the data distribution and may bring unnecessary computation in message iterative process.

The aim of this paper is to propose a new AP algorithm, Adjustable Preference Affinity Propagation (APAP) algorithm. In APAP, the value of each element preference p_k is independently determined on the basis of the data distribution in the initial stage and is automatically adjusted during the iteration process. Experiments on both synthetic data set and real-world data sets verify that the proposed APAP algorithm can obtain better performance comparing with the standard AP algorithm.

The remaining of this paper is organized as follows. Section 2 gives a brief review of AP and the preference's selection problem. Section 3 introduces the proposed APAP algorithm. Section 4 presents the experimental results on synthetic data and real data. Section 5 concludes the research work and its results.

2. AP and the related problem

2.1. AP

In 2007, Frey and Dueck published a paper entitled 'Clustering by Passing Messages between Data Points' and proposed Affinity Propagation (AP), a new exemplar-based clustering method. The distinguishing feature of AP is that it introduces the graphical model's belief propagation (max-product) algorithm into clustering, considers all data points as potential exemplars, and identifies clusters automatically.

AP considers every data point as a node ($\{1, 2, \ldots, N\}$) in a network. Let $s(i, k)$ indicate the similarity between the data point i and the data point k. AP takes a collection of real-valued similarities $\{s(i,k)\}_{N\times N}(i, k = 1, 2, \ldots, N)$ as its input and recursively transmits real-valued messages (the responsibilities $r(i, k)$ and the availabilities $a(i, k)$) along the edges until a high-quality set of exemplars and its corresponding clusters emerge. Fig. 1 shows a typical factor graph model of AP.

In Fig. 1, $s(i, c_i)$ indicates the similarity between the data point i and its corresponding exemplar c_i (in AP, c_i is a data point.). The purpose of message transmitting in AP is to search for such a set of $\{c_i\}$ that makes the global graph's function F maximized. That can be described by the following optimization problem,

$$\text{Max} : F(C; s) = e^{\sum_{i=1}^{N} s(i,c_i)+\sum_{k=1}^{N} \log f_k(C)} \quad (1)$$

Where, $C = \{c_i\}$, $f_k(C)$ is a kind of coherence constraint.

Note that the first term $(\sum_{i=1}^{N} s(i, c_i))$ in the exponent involves the net similarity S, the graph's function F can be seen as a constrained net similarity. Thus, the goal of message transmitting can be changed to make the net similarity under the coherence constraint condition maximized [2,8,15,20,22], i.e.

$$\text{Max} : S(C; s) = \sum_{i=1}^{N} s(i, c_i) + \sum_{k=1}^{N} \log f_k(C) \quad (2)$$

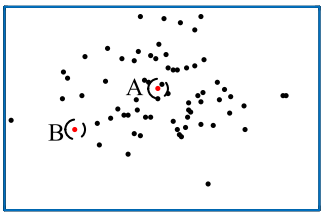

Fig. 2. A Gaussian distribution data set.

To achieve the goal, sum-product/max-product algorithm is applied. Relative iteration formulas are available in reference [2].

2.2. The problem in preference selection

Among the inputs $\{s(i,k)\}_{N\times N}(i, k = 1, 2, \ldots, N)$ of AP, there is a special parameter, $s(k, k)$, which indicates how likely the relevant data point k is to be chosen as an exemplar. $s(k, k)$ is named as the kth element preference p_k and is the element of the preference set denoted by P ($P=\{p_k\}_{N\times 1}$, $k \in \{1, 2, \cdots, N\}$, $s(k, k) = p_k$). The data point with larger value of p_k is more likely to be chosen as an exemplar [1,12].

It can be deduced from Eq. (2), the values of preference P may significantly affect AP clustering result and improper values of P may lead to suboptimal clustering solutions [14,18,29]. P can also be recognized as a control knob to govern the number of clusters [2,3,8,20,30]. The lower values of P penalize the use of data points as exemplars more heavily and lead to fewer cluster, while the effect with higher values may be opposite [2,3,14,15,21,28]. Meanwhile, many empirical results have also validated that setting P on the median of the input similarities results in a moderate number of clusters and their minimum results in a smaller number of clusters [1,2,25].

Up to date, in AP and most of AP-based algorithms, all the elements are commonly set as a same constant [1,2,4,10,13,16,18,20,21,23]. Reference [30] provided a method to compute the value of the lower bound and the upper bound of p_k, in which all the values of p_k are same and also set as a constant. However, the problem is that although every data point can be an exemplar, the probabilities may not be same in practical clustering problems. As shown in Fig. 2, it is obvious that the data point A and B have different possibilities to become an exemplar. To make all the data points share one common exemplar preference is not a good choice. That ignores the information contained in the data distribution and may cause unnecessary computation in message iterative process.

3. The proposed algorithm: adjustable preference affinity propagation (APAP) algorithm

To overcome the mentioned drawback of the standard AP and AP-based algorithms (all the values of element preferences are commonly set as a same constant), we propose a new AP algorithm, the Adjustable Preference Affinity Propagation (APAP) algorithm. The distinguishing features of APAP algorithm are that the value of each element preference p_k is set to a special value based on the data distribution in the initial stage and will be automatically adjusted according to the mutual effects among the exemplars during iteration process.

3.1. The initial values of preference $\{p_k\}_{N\times 1}$

As mentioned above, in some cases, it is not a good choice to make all the data points share one common exemplar preference.