Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 486075, 8 pages
http://dx.doi.org/10.1155/2014/486075

Research Article

An Affinity Propagation Clustering Algorithm for Mixed Numeric and Categorical Datasets

Kang Zhang and Xingsheng Gu

Key Laboratory of Advanced Control and Optimization for Chemical Processes, Ministry of Education, East China University of Science and Technology, Shanghai 200237, China

Correspondence should be addressed to Xingsheng Gu; xsgu@ecust.edu.cn

Received 5 June 2014; Accepted 4 September 2014; Published 29 September 2014

Academic Editor: Kang Li

Copyright © 2014 K. Zhang and X. Gu. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Clustering has been widely used in different fields of science, technology, social science, and so forth. In real world, numeric as well as categorical features are usually used to describe the data objects. Accordingly, many clustering methods can process datasets that are either numeric or categorical. Recently, algorithms that can handle the mixed data clustering problems have been developed. Affinity propagation (AP) algorithm is an exemplar-based clustering method which has demonstrated good performance on a wide variety of datasets. However, it has limitations on processing mixed datasets. In this paper, we propose a novel similarity measure for mixed type datasets and an adaptive AP clustering algorithm is proposed to cluster the mixed datasets. Several real world datasets are studied to evaluate the performance of the proposed algorithm. Comparisons with other clustering algorithms demonstrate that the proposed method works well not only on mixed datasets but also on pure numeric and categorical datasets.

1. Introduction

With the development of information technology and with the wide use of computer and networks, the explosion of data in almost all fields provides a totally new perspective for data scientists towards knowledge discovery and future decision. Because of the urgent need of data processing, new techniques that can extract useful information and knowledge from the vast amount of data have been developed by researchers. In this context, data mining is an effective and attractive approach to meet these requirements.

Clustering is one of the most commonly encountered data mining techniques that are implemented to extract knowledge arising from many areas, some of which are community detection [1], pattern recognition [2, 3], bioinformatics [4], and spatial database applications, for example, GIS or astronomical data [5, 6]. The general purpose of clustering is to partition a dataset consisting of n points embedded in m-dimensional space into k clusters, such that the data points within the same cluster are more similar to each other than to data points in other clusters [7–9]. Because of the simplicity and ease of implementation in a wide variety of scenarios, distance-based clustering methods, for instance, k-means, k-medians, k-medoids, and hierarchical clustering, are widely used and deeply researched. The main problems of distance-based clustering methods are defining a proper similarity measure to discriminate the similarity or dissimilarity between different data points and aggregating most similar elements into appropriate clusters in an unsupervised way. Thus, the problem of clustering can be reduced to the problem of finding a distance function for that data type [10–12]. Traditional clustering methods use Euclidean distance measure to calculate the similarity (or dissimilarity) of two data points [13, 14]. It is suitable for the datasets that are purely numeric. Actually, datasets in real world are more complicated. Large amount of data is mixed containing both numeric attributes like height, age, and so forth and categorical attributes like male or female, on or off, and so forth. In this case, however, Euclidean distance measure fails to judge the similarity of two data points when attributes are of categorical or mixed type.

Up to the present, researchers have been developing many ways dealing with mixed data. Similarity based agglomerative clustering (SBAC) [15], a hierarchical agglomerative algorithm, based on Goodall similarity measure [16], presented by Li and Biswas works well with mixed numeric and categorical

attributes. But the amount of computation, while clustering large datasets, is rapidly increasing, which is not acceptable. Huang [17] proposed k-prototype clustering method that divides the dataset into two distinct parts, one for numeric attributes and another for categorical attributes, and handles the two components separately. Due to the information loss in dealing with cluster center and the simple binary distance measure between two categorical attributes of Huang's algorithm, Ahmad and Dey [18] developed a modified cost function alleviating the shortcomings of Huang's cost function based on a k-mean type algorithm. In Ahmad and Dey's algorithm, the distance computation of two values in a single categorical attribute considers not only the attribute they belong to, but also other attributes including the numeric ones. They also proposed a significance computing approach of a numeric attribute based on the attribute value distributions within the data. Ji et al. [19] improved Ahmad and Dey's algorithm with a novel fuzzy k-prototype algorithm integrating mean and fuzzy centroid to represent the prototype of a cluster. Like many other fuzzy k-means type algorithms, Ji's algorithm also needs the determination of fuzzy coefficient value.

The novel affinity propagation clustering (APC) algorithm based on message passing is a more powerful approach proposed by Frey and Dueck [20] in 2007. Traditional distance-based clustering methods satisfy the conditions of metric similarities, that is, symmetry, nonnegativity, and the triangle inequality. Compared to the traditional approaches, the affinity propagation algorithm's ability to take as input also general nonmetric similarities makes it suitable for exploratory data analysis using unusual measures of similarity [21]. For instance, AP has been used to identify key sentences and air-travel routing on the basis of nonstandard optimization criteria [20]. Furthermore, affinity propagation is a completely data-driven analysis technique that partitions the data points to different clusters and identifies exemplars among them by simultaneously considering all data points as possible exemplars and exchanging messages between data points until a good set of exemplars and clusters emerges [22]. However, the original AP method assumes features are numeric valued, which means the algorithm cannot process features of categorical values or mixed type values.

Based on AP algorithm and Ahmad and Dey's mixed similarities measure architecture [18], this paper proposes an adaption affinity propagation clustering method for mixed numeric and categorical attributes datasets using a novel similarity measure as a cost function. The key innovative points of the paper are as follows. (1) This paper has applied the AP algorithm to cluster mixed type attributes datasets for the first time. (2) This paper proposes a novel mixed similarities measure based on Ahmad and Dey's work. (3) The method improves the original AP clustering algorithm with adaption strategies.

The rest of the paper is organized as follows. We start in Section 2 with a brief review of affinity propagation clustering algorithm and the distance measure for mixed type datasets. In Section 3, the novel similarity measure for mixed type data is introduced and then the novel adapted AP approach is described in detail. Section 4 presents the experimental methodology and results on several benchmark datasets as well as the comparisons with the selected baseline algorithms. Discussions and conclusions are given in Section 5.

2. Background

2.1. Description of AP. Exemplar-based clustering, such as the popular k-centers and k-medians clustering methods, partitions the dataset by identifying a subset of representative elements (exemplars), so that the sum of distances between data points and their exemplars is minimized [23]. The traditional clustering analysis methods usually start with an initialization step that the algorithm selects K initial data centers as exemplars and allocates other data points based on the distances to exemplars. It is obvious that different initial selection comes to different clustering results. On the contrary, AP runs based on an entirely different mechanism. Firstly, all data points are considered as potential exemplars and they are viewed as nodes in a network. Secondly, a number of real-valued messages are iteratively transmitted along edges of the network so that a relevant set of exemplars and corresponding clusters is identified [20]. Details of the framework are as follows [24].

AP takes as input a matrix of real-valued similarities between data points. Let $\{s_{ij}\}$, $i = 1, \ldots, N$, $j = 1, \ldots, N$, be a set of N^2 real-valued variables where s_{ij} indicates the similarity between two objects x_i and x_j in it. AP defines s_{ij} as the negative of the square of their Euclidean distance; that is, $s(i, j) = -\|x_i - x_j\|^2$, $i \neq j$. The self-similarities s_{kk} are referred to as "preferences" that influence the probability of one point being an exemplar. If there is no a priori knowledge, preferences are set to common values so that each data point is regarded as a potential exemplar with equal probability.

As mentioned above, in AP algorithm, data points exchange information by passing messages. Two kinds of messages are produced in the procedure and each takes into account a different kind of competition. One is called"‘responsibility" $r(i, k)$, which is sent from data point i to candidate exemplar point k. $r(i, k)$ reflects the accumulated evidence for how well suited point k is to serve as the exemplar for point i, taking other potential exemplars for point i into consideration. The "responsibility" $r(i, k)$ is updated as follows:

$$r(i,k) \longleftarrow s(i,k) - \max_{k' \text{ s.t. } k' \neq k} \left\{ a\left(i,k'\right) + s\left(i,k'\right) \right\}. \quad (1)$$

The other one is called "availability" $a(i, k)$, gathering evidence from data points as to whether each candidate exemplar would make a good exemplar. It is sent from the candidate representative point k to point i, reflecting the accumulated evidence for how appropriate it would be for point i to choose point k as its exemplar. Beside, the support from other points that point k should be an exemplar is taken into account. The "availability" $a(i, k)$ is updated as follows:

$$a(i,k) \longleftarrow \min \left\{ 0, r(k,k) + \sum_{i' \text{ s.t. } i' \notin \{i,k\}} \max \left\{ 0, r\left(i',k\right) \right\} \right\}. \quad (2)$$

The "self-availability" $a(k,k)$ reflects accumulated evidence that point k is an exemplar, based on the positive responsibilities sent to candidate exemplar k from other points. $a(k,k)$ is updated differently as follows:

$$a(k,k) \longleftarrow \sum_{i' \text{ s.t. } i' \neq k} \max\left\{0, r\left(i', k\right)\right\}. \tag{3}$$

After iterative message passing, exemplars can be identified by calculating maximum of $a(i,k) + r(i,k)$ for point i. If $k = i$, point i is selected as an exemplar, or point k is the exemplar of point i.

Furthermore, in order to avoid numerical oscillations in some circumstances when updating the messages, the damping factor λ is introduced to iteration process:

$$\begin{aligned} r_{t+1}(i,k) &= \lambda \cdot r_t(i,k) + (1-\lambda) \cdot r_{t+1}(i,k), \\ a_{t+1}(i,k) &= \lambda \cdot a_t(i,k) + (1-\lambda) \cdot a_{t+1}(i,k), \end{aligned} \tag{4}$$

where t indicates the iteration times.

The primary advantage of AP algorithm is that AP does not need to preassign the number of clusters, which is different with k-means methods specifying the K value. This is because AP considers each data point as a potential exemplar and the probability of being an exemplar depends on the shared value of preference. With greater value of preference, AP generates more clusters. Another advantage is that AP only accepts the collection of similarities as input, which eliminates the need to deal with the raw dataset directly. This is an instrumental feature with which researchers can determine the similarity input matrix using various distance measures that are suitable for the objects of clustering. Moreover, wide-ranging applications [20] of AP demonstrate its ability of processing large datasets rapidly and effectively.

However, AP meets some limitations as well. The specific value of "preferences," for the procedure of clustering, is a double-edged sword. Frey and Dueck [20] suggested setting the shared value of preferences as the median of the input similarities resulting in a moderate number of clusters. It is difficult to determine suitable value of preference in which different values lead to completely different results if there is no a priori knowledge. In addition, the damping factor λ acquires an appropriate value. Equation (4) shows that a larger value of damping factor means not easily trapping into oscillations but reducing the convergence rate, while a smaller one results in a fast rate of convergence but with a risk of no convergence when the message-passing procedure is terminated. Frey and Dueck [20] suggested a default damping factor of $\lambda = 0.5$ to keep balance between convergence and oscillation.

2.2. Distance and Significance. Based on Huang's cost function [17], Ahmad and Dey developed a motivated distance and significance computation framework that not only considers the distances between pairs of distinct values within an attribute, but also takes their cooccurrence with other attributes into account [18]. Two parts are introduced to generate the distance matrix of a mixed dataset.

The first step is to calculate the distance between each pair of values for categorical attributes. For the given mixed type dataset, A_i denote a categorical attribute, in which x and y are two of the values. Let A_j denote another categorical attribute and w a subset of values of A_j. Accordingly, $\sim w$ denotes the complementary set of w. The conditional probability $P_i(w \mid x)$ means the probability that a data point having value x for A_i has a value belonging to w for A_j. Likewise, $P_i(\sim w \mid y)$ denotes the conditional probability that a data point having value y for A_i has a value belonging to $\sim w$ for A_j.

The distance between the pair of values x and y of A_i as regards the attribute A_j and a particular subset w is defined as follows:

$$\delta_w^i(x,y) = P_i(w \mid x) + P_i(\sim w \mid y). \tag{5}$$

Distance between attribute values x and y for A_i concerning attribute A_j is denoted by $\delta^{ij}(x,y)$ and is given by

$$\delta^{ij}(x,y) = P_i(\omega \mid x) + P_i(\sim \omega \mid y), \tag{6}$$

where ω is the subset w of values of A_i that maximizes the quantity $P_i(\omega \mid x) + P_i(\sim \omega \mid y)$. Considering both $P_i(\omega \mid x)$ and $P_i(\sim \omega \mid y)$ lie between 0 and 1.0, in order to restrict the value of $\delta^{ij}(x,y)$ between 0 and 1.0, $\delta^{ij}(x,y)$ is defined as

$$\delta^{ij}(x,y) = P_i(\omega \mid x) + P_i(\sim \omega \mid y) - 1.0. \tag{7}$$

For a dataset with m attributes, including categorical and numeric attributes which have been discretized, the distance between two dissimilar values x and y of any categorical attribute is given by

$$\delta(x,y) = \frac{\sum_{j=1,\ldots,m, i \neq j} \delta^{ij}(x,y)}{m-1}. \tag{8}$$

Using (5) to (8), it is possible to compute the distance between two distinct values of categorical attributes and the discretized numeric attributes.

In the second step, significance for each numeric attribute is determined. Significance of an attribute defines the importance of that attribute in the dataset [25, 26]. To compute the significance of a numeric attribute, it must be discretized to S intervals first. Thus each interval can be assigned a categorical value $c[1], c[2], \ldots, c[S]$. Therefore, using (5) to (8), in the same way as it is computed for categorical values, the distance $\delta(c[i], c[j])$ for every pair of discretized numeric values $c[i]$ and $c[j]$ can be computed. Eventually, the significance of a numeric attribute, ω_r, is computed as the mean of $\delta(c[i], c[j])$ for all pairs $c[i] \neq c[j]$:

$$\omega_r = \frac{\sum_{i=1}^{S} \sum_{j>i}^{S} \delta\left(c[i], c[j]\right)}{S(S-1)/2}. \tag{9}$$

3. Method

The main idea of our proposed algorithm is that our method attempts to obtain a clustering result using a similarity measure for mixed data based on AP clustering approach. We propose the novel method to handle the problem in the following sections.

Set $i = 0$;
for each numeric attribute A_i in dataset A **do**
  Figure out the similarity matrix based on (10) as the input;
  Calculate the median of similarities as the shared value of preference;
  Perform the AP algorithm using (1)–(4) to obtain an S_i classification result;
  Discretize attribute A_i to S_i intervals according to the clustering result;
  $i = i + 1$;
end for
Establish a new dataset B which is a pure categorical dataset composed of the discretized numeric attributes and the original categorical attributes;
for each attribute B_j in dataset B **do**
  Calculate the distance between two distinct values of any categorical attribute using (5)–(8);
  Compute the significance (weight) of each numeric attribute using (9) in which the interval S is replaced by S_i.
end for

Algorithm 1: Pseudocode of computing significance.

3.1. Improved Similarity Measure. The main advantage of Ahmad and Dey's distance measure method is that it considers the distance between a pair of values for an attribute as the function of their cooccurrence probabilities with a set of values of another categorical attribute. Therefore, the distance is a reflection of the difference between two categorical values rather than being the same or different (0 or 1). On the other hand, significance of an attribute is not user-defined like other algorithms but computed based on the discretization of numeric attributes. In other words, the algorithm, by itself, decides which attribute is more important and assigns a higher weight to it.

However, the distance measure also faces some limitations. In the process of discretization of numeric attributes, a number of the intervals (S) of numeric values are user-defined referring to different problems, equally for all numeric attributes, which will cause an inaccurate discretization because the algorithm has no consideration on the different distribution of distinct attributes. Beside, one should test the parameter S to find a suitable interval for discretizing numeric values.

As mentioned in Section 2.1, AP algorithm separates data objects into suitable clusters without assigning the object number of classes, since each data point is viewed as a potential exemplar. Therefore, we propose an improved similarity measure based on Ahmad and Dey's work in which the discretization operation is replaced by AP clustering discretization approach. Data objects are allocated to clusters as natural as they are distributed. Furthermore, different intervals emphasize the distinction of each attribute influencing the significance values.

For a given mixed dataset, let A_i, $i = 1, \ldots, r$, denote a numeric attribute, whose values are $\{x_1, x_2, \ldots, x_n\}$, where r is the number of numeric attributes and n is the number of data points. The similarity between x_1 and x_2 is defined by

$$s(x_1, x_2) = -(x_1 - x_2)^2, \tag{10}$$

where s can be viewed as an $n \times n$ matrix and the similarity $s(x_k, x_j)$ indicates the negative squared error for points x_k and x_j. The novel method for computing significance of numeric attribute is listed in Algorithm 1.

Figure 1 illustrates the performance of three different discretization techniques. Raw data, equal width, equal frequency, and AP method are listed in the figure. It shows that AP method performs the best reflection for the distribution regularity of the data points.

Let S denote the $N \times N$ similarity matrix, and we define the similarity between two values x_i and x_j as follows:

$$S(x_i, x_j) = -\sum_{t=1}^{m_r} \left(\omega_t \cdot (x_{it}^r - x_{jt}^r)\right)^2 - \sum_{t=1}^{m_c} \left(\delta\left(x_{it}^c, x_{jt}^c\right)\right)^2, \tag{11}$$

where $\sum_{t=1}^{m_r} (\omega_t \cdot (x_{it}^r - x_{jt}^r))^2$ denotes the distance of objects x_i and x_j for numeric attributes only, ω_t is the significance of the ith numeric attribute described in Section 3.1, and $\sum_{t=1}^{m_c} (\delta(x_{it}^c, x_{jt}^c))^2$ denotes the distance between data objects x_i and x_j in terms of categorical attributes only. The similarities are set to a negative squared error to coordinate the input of AP algorithm.

3.2. Adaptive AP Algorithm. In Section 2.1, we discussed the advantages and limitations of AP algorithm. The shared value of "preferences" (p) is the key value that determines the clustering performance as well as the number of classes in the result. In some cases, the objective number of clusters is preassigned while it is hard to define an appropriate value of p. This is because there is not a one-to-one correspondence between the output number of classes and the value of p which means a certain range of values will arrive at the same number of clusters, with different distributions yet. To search the optimal p value for the given number of classes, an adaptive strategy is proposed as follows:

$$p_{t+1} = \alpha p_t + \beta, \tag{12}$$

where t denotes the running time of AP algorithm, $\alpha > 1$ is a function of the number of clusters K_t in the tth running

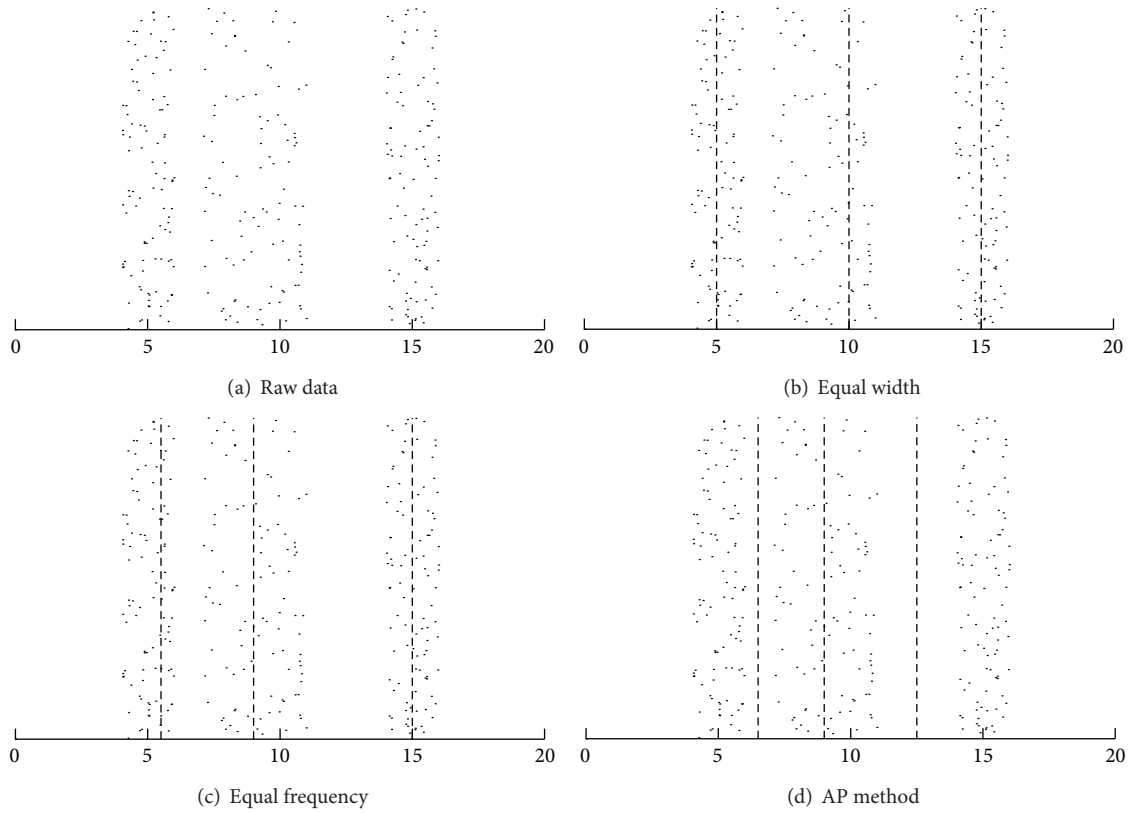


Figure 1: Different discretization techniques. (a) The raw data are assigned a random y-axis component such that the data points can be distinguished. (b) The data points are divided by equal width discretization method that cannot work well obviously. (c) The data points are divided by equal frequency method that is better than equal width method, but not the best. (d) The data points are well divided by AP method we proposed, and it is the best one of the three.

result, and β denotes the searching step. p_{t+1} and p_t are negative values.

We named α as coarse tuning coefficient while β was named as fine tuning coefficient. When the value of K_t is much greater than the target value K, a relevant greater value of α should be employed, so that the p value may reduce quickly. On the contrary, when K_t is close to the target K, smaller value of α should be defined. So we set the coarse tuning coefficient as $\alpha = \sqrt{(K_t - K) + 0.5}$. In this case, the algorithm is able to tune the value of α dynamically, according to the current cluster number K_t.

Since the coarse tuning strategy makes the algorithm obtain the right number of clusters, fine tuning steps lead to the better clustering performance. In the iteration stage of $K_t \neq K$, β is set to 0. Meanwhile, when entering the stage of $K_t = K$, α is assigned to 1. Value of β is important for scanning local area to maximize the energy function. Referring to the settings in [27], β is defined as $\beta = 0.01 p_m$, where p_m denotes the initial value of p. The scanning stage may be terminated after the energy function decreases or after a fixed number of iterations of fine tuning.

On the other hand, the damping factor λ is another parameter that controls the convergence and the speed of algorithm. Our intention is that, in the case of no oscillation, the algorithm is able to acquire a faster convergence speed. An adaptive mechanism of λ is adopted to balance the contradiction between oscillation and convergence.

Although maintaining a larger λ close to 1 may avoid numerical oscillations much more easily, a homologous decline of the updating rate for "availability" and "responsibility" becomes inevitable. The algorithm needs more iteration times than that with smaller λ to obtain a corresponding result. Therefore, a changing λ along with the iteration of algorithm is a better choice. According to this conception, we have designed an adaptive mechanism for λ as follows:

$$\lambda_K = -\left(\frac{K}{\text{iteration}}\right)^a \cdot (\lambda_{\max} - \lambda_{\min}) + \lambda_{\max}, \quad (13)$$

where K denotes current number of iterations and iteration denotes maximum iteration. $\lambda_{\max}$ and $\lambda_{\min}$ denote the maximum and minimum values of λ, respectively. We introduce the coefficient a to adjust the rate of descent for λ. When the value of a is greater than 1, λ declines from flat to sharp. We recommend a to be greater than 1 to guarantee a smooth iteration process.

3.3. The Proposed Algorithm. Based on the above explanations, the pseudocode of the proposed algorithm is listed in Algorithm 2.

Calculate the similarity matrix and preferences as input of AP algorithm.
for each numeric attribute A_i in dataset A **do**
  Calculate the significance of attribute A_i using the method in Section 3.1;
end for
for each categorical attribute A_j in dataset A **do**
  Calculate the distance of any pairs of values in A_j based on (5)–(8);
end for
Generate the input similarity matrix of the mixed dataset using (11);
Set the value preference p_m by the median of similarities;
Input the target number of clusters as K;
while the termination conditions are not met **do**
  for each running time t of AP algorithm **do**
    if $K_t \neq K$ **then**
      The p value adaptive strategy is defined as $p_{t+1} = \alpha p_t$, where $\alpha = \sqrt{(K_t - K) + 0.5}$;
    else if $K_t = K$ **then**
      The p value adaptive strategy is defined as $p_{t+1} = p_t + \beta$, where $\beta = 0.01 p_m$;
    end if
    λ adaptive strategy is defined by (13);
  end for
end while

Algorithm 2: Pseudocode of the proposed algorithm.

4. Experimental Evaluation

In this section, experimental results are presented by our proposed approach and other popular clustering methods on several standard datasets. To validate the effectiveness of the proposed algorithm, we have chosen four different kinds of datasets: pure numeric (Iris), pure categorical (Zoo), and mixed type (Heart diseases and Credit Approval), which are taken from UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/index.html).

Four well-performed clustering methods including k-prototype [17], SABC [15], Ahmad and Dey's algorithm [18], and fuzzy k-prototype [19] are employed for the comparison. All clustering results have been obtained with random initialization if needed. The experiments were carried out on a workstation with a 3.40 GHz Intel i7-2600 CPU and 4.0 GB RAM. All the programs were written by C code complying with GCC 4.6 and the operating system was Fedora Linux.

To evaluate the performance of clustering method, micro-precision [28] is introduced. Let a_i denote the number of data points that are correctly assigned to the class C_i, let b_i denote the data objects that are incorrectly assigned to the class C_i, and let c_i denote the data objects that are incorrectly rejected from the class C_i. The precision of the ith class is defined as

$$p_i = \frac{a_i}{a_i + b_i}, \tag{14}$$

and recall is defined as

$$r_i = \frac{a_i}{a_i + c_i}. \tag{15}$$

And microprecision (micro-p) is defined as

$$\text{micro precision} = \frac{\sum_{i=1}^{C} a_i}{n}, \tag{16}$$

where n is number of data objects in dataset and C denotes the number of classes for a given clustering. According to this evaluation, a higher value of micro-p indicates a better clustering result.

Table 1: Comparative study of significance of attributes for Iris data.

Attribute of Iris data	A_1	A_2	A_3	A_4
Significance [18]	0.70	0.67	0.78	0.77
Significance (proposed algorithm)	0.52	0.49	0.68	0.69

4.1. Iris. Iris dataset consists of 150 instances of 3 classes: Iris Setosa, Iris Versicolour, and Iris Virginica. All 4 attributes are pure numeric in this dataset. Because there are no categorical attributes in the dataset, we just need to discretize the numeric values and calculate the significance of each attribute. Here, a comparison of significance between [18] and our proposed algorithm is illustrated in Table 1. From this table, we can see that each attribute obtains different weight value by the two significance calculation algorithms. By [18]'s algorithm, the biggest difference value is 0.11 between 0.67 of A_2 and 0.78 of A_3. By our proposed algorithm, the biggest difference value is 0.20 between 0.49 of A_2 and 0.69 of A_4. It is obvious that more significant difference of weight was calculated by the proposed algorithm which helps in the discrimination of different attributes.

A comparison of clustering results is presented in Table 2. Ahmad and Dey's algorithm, k-mean, original AP, and our proposed algorithm give the microprecisions 0.947, 0.88, 0.9, and 0.947, respectively. In this case, we can see that the proposed algorithm obtains a better result than the original

Table 2: Comparative study of different clustering algorithms for Iris dataset.

Algorithm	Number of data objects in desired clusters	Micro-p
Ahmad and Dey's algorithm	142	0.947
k-mean	132	0.880
AP	135	0.900
Proposed algorithm	142	0.947

AP algorithm due to the introduction of the significance value of each attribute and the improvements based on the original AP algorithm. Because of the simplicity of Iris dataset, the proposed method does not show significant advantages to Ahmad and Dey's algorithm.

4.2. Zoo. Zoo dataset contains 101 instances distributed into 7 classes. Since all 16 attributes are of categorical type, only those algorithms that can deal with categorical data or mixed type data are applied in this section. For the proposed algorithm, only categorical part remains in the similarity measure function 11. Fuzzy k-prototype algorithm uses the same similarity measure as Ahmad and Dey's algorithm. Table 3 shows the microprecision values of each clustering algorithm on Zoo dataset. K-prototype, SABC, fuzzy k-prototype, and our proposed algorithm give the microprecisions 0.802, 0.426, 0.908, and 0.921, respectively. From the table, we can see that the proposed algorithm allocated 93 data objects in desired clusters from the total 101 instances while the other three algorithms give the numbers 81, 43, and 91, respectively.

4.3. Heart Diseases. Heart diseases database contains 76 attributes, and 14 of them are for the experiments. The 14th attribute is the class index. This dataset is a mixed one with 8 categorical and 5 numeric features. 303 instances belong to 5 classes, that is, 1 for normal (164) and 4 for sick (139). Table 4 shows the results of 4 clustering methods. K-prototype, SABC, fuzzy k-prototype, and our proposed algorithm give the microprecisions 0.545, 0.545, 0.717, and 0.886, respectively. From the table, we can see that the proposed algorithm allocated 263 data objects in desired clusters from the total 303 instances while the other three algorithms give the numbers 162, 162, and 213, respectively. The proposed algorithm performs better than the other three algorithms.

4.4. Credit Approval. Credit Approval dataset, from credit card organization, is also a mixed dataset containing 9 categorical and 6 numeric attributes. 690 instances of data objects are divided into 2 classes: positive (307) and negative (383). Table 5 presents the results of 5 clustering algorithms. K-prototype, SABC, Ahmad and Dey's algorithm, fuzzy k-prototype, and our proposed algorithm give the microprecision values 0.562, 0.555, 0.883, 0.838, and 0.920, respectively. Ahmad and Dey's algorithm and fuzzy k-prototype use the same similarity measure so they get close values in the result. 635 data objects from the total 690 instances were clustered in their desired clusters while the other four algorithms give 388, 383, 609, and 578, respectively. Our approach achieves better result in the comparison.

Table 3: Comparative study of different clustering algorithms for Zoo dataset.

Algorithm	Number of data objects in desired clusters	Micro-p
K-prototype	81	0.802
SABC	43	0.426
Fuzzy k-prototype	91	0.908
Proposed algorithm	93	0.921

Table 4: Comparative study of different clustering algorithms for Heart diseases dataset.

Algorithm	Number of data objects in desired clusters	Micro-p
K-prototype	162	0.545
SABC	162	0.545
Fuzzy k-prototype	213	0.717
Proposed algorithm	263	0.886

Table 5: Comparative study of different clustering algorithms for Credit Approval dataset.

Algorithm	Number of data objects in desired clusters	Micro-p
K-prototype	388	0.562
SABC	383	0.555
Ahmad and Dey's algorithm	609	0.883
Fuzzy k-prototype	578	0.838
Proposed algorithm	635	0.920

5. Conclusion

Extracting knowledge and information from mixed data meets the urgent needs of real world applications. Affinity propagation is a novel unsupervised clustering method presented in recent years. In this paper, we proposed a new approach for clustering mixed numeric and categorical data based on AP method. We made the contribution of three aspects. Firstly, we extend AP method to deal with the mixed type dataset removing its numeric data limitation and the results have shown the feasibility of this extension. Secondly, an improved mixed similarity measure is proposed to compute distances between pairs of values for categorical attribute and to obtain the weight coefficients for numeric attribute. Finally, we improve the original AP by employing adaption strategies.

Our approach works well not only for mixed data clustering but also for clustering pure numeric or categorical data,

which has been demonstrated in the experiments by comparing with other clustering algorithms. The experimental results illustrate the efficiency of the proposed method on several real life mixed type datasets. However, like many other algorithms with parameter tuning problem, we introduce several user-defined parameters, and it is not always clear which is the best value for these parameters. Our future work will focus on the further improvement of AP algorithm and its applications on various fields.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors are grateful to the support of the National Natural Science Foundation of China (Grant nos. 61174040, 61104178, and 61205017), Shanghai Commission of Science and Technology (Grant no. 12JC1403400), and the Fundamental Research Funds for the Central Universities.

References

[1] S. Fortunato, "Community detection in graphs," *Physics Reports A: Review Section of Physics Letters*, vol. 486, no. 3–5, pp. 75–174, 2010.

[2] A. Baraldi and P. Blonda, "A survey of fuzzy clustering algorithms for pattern recognition, I," *IEEE Transactions on Systems, Man, and Cybernetics B: Cybernetics*, vol. 29, no. 6, pp. 778–785, 1999.

[3] A. Baraldi and P. Blonda, "A survey of fuzzy clustering algorithms for pattern recognition—part II," *IEEE Transactions on Systems, Man, and Cybernetics B: Cybernetics*, vol. 29, no. 6, pp. 786–801, 1999.

[4] D. Tang, Q. Zhu, and F. Yang, "A Poisson-based adaptive affinity propagation clustering for SAGE data," *Computational Biology and Chemistry*, vol. 34, no. 1, pp. 63–70, 2010.

[5] M. Ester, A. Frommelt, H.-P. Kriegel, and J. Sander, "Spatial data mining: database primitives, algorithms and efficient DBMS support," *Data Mining and Knowledge Discovery*, vol. 4, no. 2-3, pp. 193–216, 2000.

[6] J. Sander, M. Ester, H.-P. Kriegel, and X. Xu, "Density-based clustering in spatial databases: the algorithm gdbscan and its applications," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 169–194, 1998.

[7] J. Han, M. Kamber, and J. Pei, *Data Mining*, The Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann, Boston, Mass, USA, 3rd edition, 2012.

[8] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264–323, 1999.

[9] P. Berkhin, "A survey of clustering data mining techniques," in *Grouping Multidimensional Data*, pp. 25–71, Springer, Berlin, Germany, 2006.

[10] C. C. Aggarwal, "Towards systematic design of distance functions for data mining applications," in *Proceeding of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '03)*, pp. 9–18, New York, NY, USA, August 2003.

[11] B. McCane and M. Albert, "Distance functions for categorical and mixed variables," *Pattern Recognition Letters*, vol. 29, no. 7, pp. 986–993, 2008.

[12] C. C. Aggarwal and C. K. Reddy, *Data Clustering: Algorithms and Applications*, CRC Press, 2013.

[13] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, vol. 3, Wiley, New York, NY, USA, 1973.

[14] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*, Prentice Hall, Inc., Upper Saddle River, NJ, USA, 1988.

[15] C. Li and G. Biswas, "Unsupervised learning with mixed numeric and nominal data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 4, pp. 673–690, 2002.

[16] D. W. Goodall, "A new similarity index based on probability," *Biometrics*, vol. 22, no. 4, pp. 882–907, 1966.

[17] Z. X. Huang, "Clustering large data sets with mixed numeric and categorical values," in *Proceedings of the 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD '97)*, pp. 21–34, Singapore, 1997.

[18] A. Ahmad and L. Dey, "A k-mean clustering algorithm for mixed numeric and categorical data," *Data and Knowledge Engineering*, vol. 63, no. 2, pp. 503–527, 2007.

[19] J. Ji, W. Pang, C. Zhou, X. Han, and Z. Wang, "A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data," *Knowledge-Based Systems*, vol. 30, pp. 129–135, 2012.

[20] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, 2007.

[21] R. Guan, X. Shi, M. Marchese, C. Yang, and Y. Liang, "Text clustering with seeds affinity propagation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 4, pp. 627–637, 2011.

[22] J. Zhang, X. Tuo, Z. Yuan, W. Liao, and H. Chen, "Analysis of fMRI data using an integrated principal component analysis and supervised affinity propagation clustering approach," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 11, pp. 3184–3196, 2011.

[23] I. E. Givoni, C. Chung, and B. J. Frey, "Hierarchical a finity propagation," CoRR abs/1202.3722.

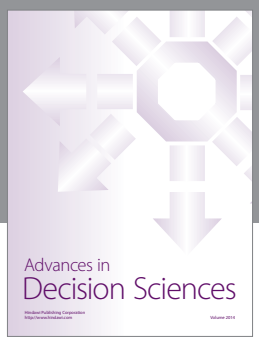
[24] I. E. Givoni and B. J. Frey, "A binary variable model for affinity propagation," *Neural Computation*, vol. 21, no. 6, pp. 1589–1600, 2009.

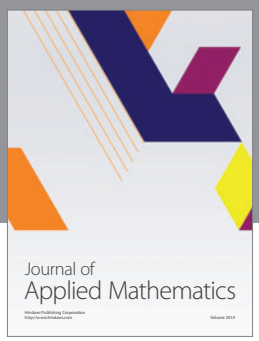
[25] J. Basak, R. K. De, and S. K. Pal, "Unsupervised feature selection using a neuro-fuzzy approach," *Pattern Recognition Letters*, vol. 19, no. 11, pp. 997–1006, 1998.

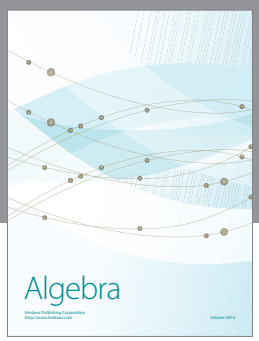
[26] D. S. Yeung and X. Z. Wang, "Improving performance of similarity-based clustering by feature weight learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 556–561, 2002.

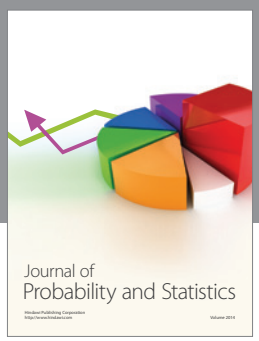
[27] K. Wang, J. Zhang, D. Li, X. Zhang, and T. Guo, "Adaptive a finity propagation clustering," CoRR abs/0805.1096.

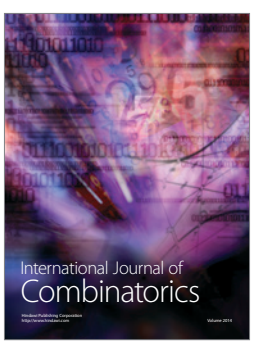
[28] Y. Yang, "An evaluation of statistical approaches to text categorization," *Information Retrieval*, vol. 1, no. 1-2, pp. 69–90, 1999.

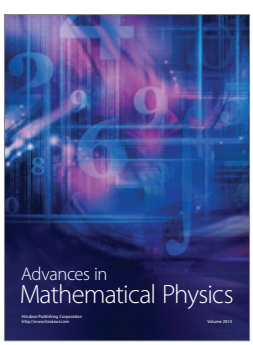


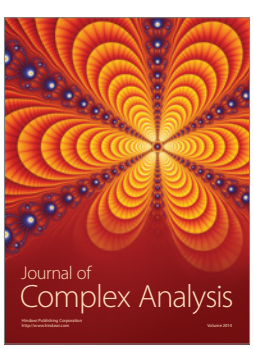


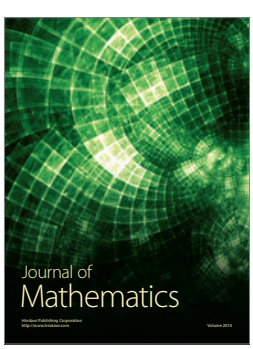


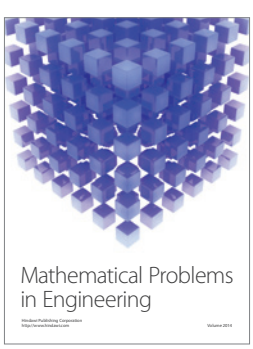


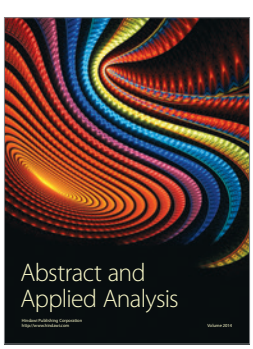


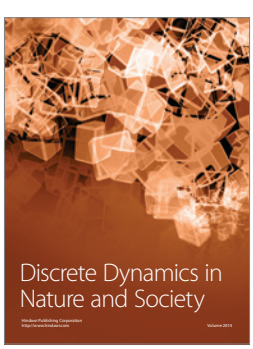


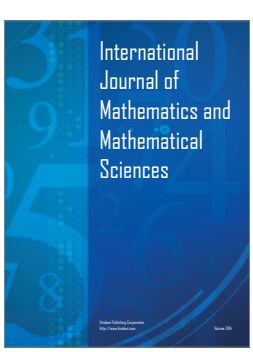












Hindawi

Submit your manuscripts at
http://www.hindawi.com