Adaptive Clustering Algorithm Based on Max-min Distance and Bayesian Decision Theory

Fengqin Zhao, Youlong Yang, Weiwei Zhao

***Abstract*—K-means clustering algorithm is one of the most famous partitioning clustering techniques that have been widely applied in many fields. Although it is very simple and fast in the process of clustering, the method suffers from a few drawbacks. K-means clustering algorithm requires to specifying the number of clusters which is difficult to know in advance for many real data sets. In addition, K-means clustering algorithm often leads to different clustering results because initial seeds are chosen randomly. To solve these problems, this paper proposes an adaptive clustering algorithm. The new algorithm adopts the idea of continuous partition of a given data set. In the process of each partition, the algorithm can select initial seeds based on max-min distance to obtain a certain result of clustering, and it can evaluate the risk of the clustering result by extending Bayesian decision theory to the field of clustering. Comparing the risk values before and after partitioning, the algorithm can decide whether the data set is continue partitioned, thus it can determine the number of clusters and get the final result of clustering automatically. The performance of the proposed algorithm has been studied on some synthetic and real world data sets. The experimental results illustrate that the new algorithm, without parameter specified by users in advance, is able to obtain efficient clustering results.**

***Index Terms*—Adaptive clustering algorithm, K-means clustering algorithm, Bayesian decision theory, Max-min distance.**

I. Introduction

CLUSTERING is one of the most efficient data mining tools to reveal intrinsic structures in a data set [1]. In general, clustering categorizes a set of unlabeled objects into some subsets (called clusters) such that the similarities among objects belonging to the same cluster are larger than the similarities among objects belonging to different clusters. As clustering does not require any information other than the given data set, it has been widely applied in many fields such as wireless sensor networks [2], image analysis [3], pattern recognition [4], recommender systems [5], information retrieval [6], [7], bioinformatics [8] and so on.

With the continuous research of clustering technique, a large number of clustering algorithms have been proposed in the literature. Generally speaking, these algorithms can be classified as hierarchical clustering [9], [10], partitioning clustering [9], [11], density-based clustering [12], [13], grid-based clustering [14], [15] and so on. Among them, partitioning clustering algorithms is applied widely because of its rapidity and effectiveness.

Manuscript received September 08, 2016; revised January 06, 2017. This work was supported by the National Natural Science Foundation of China (Grant No. 61573266).

Fengqin Zhao is with the School of Mathematics and Statistics, Xidian University, Xi'an, Shaanxi 710126, China. Tel:+86 18302938973. E-mail: zhaofengqintg@126.com

Youlong Yang and Weiwei Zhao are with the School of Mathematics and Statistics, Xidian University, Xi'an, Shaanxi 710126, China.

K-means is one of the most classical and well-researched examples of partitioning clustering algorithms. It was first proposed by MacQueen in 1967 [16]. In 1982, Stuart Lloyd proposed a simple and effective statistical clustering technique—K-means clustering algorithm [17]. K-means clustering algorithm needs users input on the number of clusters. It chooses the same number of initial seeds randomly as the number of clusters input by users in advance. Each data point is then assigned to the initial seed that has the minimum distance with the data point. Thus the database is initially grouped into some clusters. Once the data points are grouped to the seeds a new seed for each cluster is again computed. Then the database is partitioned again according to the new seeds. The above two steps are repeated until the seeds do not change anymore. K-means clustering algorithm is used widely due to its simplicity and rapidity.

Despite being used in many areas, the K-means clustering algorithm is not exempt from drawbacks. On the one hand, the K-means clustering algorithm is assumed that the value of K is known by users in advance. Unfortunately, the number of clusters is hard to obtain in practice ordinarily. On the other hand, although the convergence of the K-means clustering algorithm has been proved [18], there is no guarantee of achieving a certain result of clustering because initial seeds are chosen randomly.

To combat the above shortcomings of K-means clustering algorithm, a lot of efforts have been made by researchers. As for the choice of the initial seeds, researchers have put forward many improved methods. Al-Daoud et al. [19] proposed a new method for the initialization of clusters. In 2007, Arthur et al. [20] presented a method called K-means++ which can avoid poor clustering results produced by the K-means clustering algorithm in some extent. Furthermore, there are some solutions based on the density of data points [21], [22], [23]. Some improved methods based on optimization algorithm [24], [25] and genetic algorithm [26], [27] have also been proposed to get better initial seeds. In order to determining the number of clusters, many scholars have presented many methods from different previews. Yu et al. [28] proposed a method called FACA-DTRS algorithm to determine the number of clusters using decision-theoretic rough set in 2014. A method called best K-means [29] is an improved K-means clustering algorithm, which can automatically determine the number of clusters. Furthermore, Rezaee et al. [30] has proved that the optimal number of clustering is in between 2 to $\sqrt{n}$, where n is the number of all the data points in data space.

In order to perfect the previous work and to solve the problem for determining the result of clustering without manual parameters, this paper proposes an adaptive clustering algorithm. Firstly, Bayesian decision theory is extended to the field of clustering as well as a risk assessment function for

clustering scheme is constructed. Then, the method to select the initial seeds for K-means based on max-min distance is proposed. According to the content mentioned above, the adaptive clustering algorithm is proposed in the end. The adaptive clustering algorithm adopts the idea of continuous partition of a given data set. In the process of each partition, the algorithm can select initial seeds based on max-min distance to obtain a certain result of clustering, and it can evaluate the risk of the clustering result by the proposed risk assessment function. Comparing the risk values before and after partitioning, the algorithm can decide whether the data set is continue partitioned, thus it can determine the number of clusters and get the final result of clustering automatically. Some experiments conducted on some synthetic and real world data sets illustrates that the proposed method, without manual parameters, can get reasonable clustering result.

The rest of the paper is organized as follow. In Section 2, some basic theories are reviewed. Details of the proposed algorithm are introduced in Section 3. Section 4 present experimental results and the corresponding analysis. The paper ends with conclusions and further research topics in Section 5.

II. Basic theory

In this section, we will review the generic framework of K-means clustering algorithm and the basic procedure of Bayesian decision theory.

A. K-means clustering algorithm

The K-means algorithm is one of the most widely used clustering algorithms, and it is very simple and easy to understand. The main idea of the algorithm has been introduced in the introduction. Here, we only describe the procedure of K-means as shown in algorithm 1.

Algorithm 1 K-means clustering algorithm

Input: data set $X = \{x_1, x_2, \cdots, x_n\}$, the number of clusters K.

Output: the result of clustering (clustering scheme).

1 Select K initial cluster centers $c_1, c_2, \cdots, c_K$ randomly from the given n points $\{x_1, x_2, \cdots, x_n\}, K \leq n$.

2 Assign each point x_i, $i = 1, 2, \cdots, n$ to the cluster C_j corresponding to the cluster center c_j, for $j = 1, 2, \cdots, K$ iff $\|x_i - c_j\| \leq \|x_i - c_p\|$, $p = 1, 2, \cdots, K$ and $j \neq p$.

3 Compute new cluster centers $c_1^*, c_2^*, \cdots, c_K^*$ as follows $c_i^* = \frac{1}{n_i} \sum_{x_i \in C_i} x_i$ for $i = 1, 2, \cdots, K$. where n_i is the number of data points belonging to the cluster C_i.

4 If $c_i^* = c_i$, $\forall i = 1, 2, \cdots, K$, then terminate. Otherwise, let $c_i = c_i^*$, continue from step 2.

B. Bayesian decision theory

Bayesian decision theory is a basic statistical approach that is based on quantifying the tradeoffs among various decisions using probability and the costs that accompany such decisions. In this subsection, we will describe Bayesian decision theory briefly and give an example to illustrate the procedure of Bayesian decision.

TABLE I
ALL THE VALUES OF LOSS FUNCTIONS

Actions \ States	s_1	s_2	$\cdots$	s_j	$\cdots$	s_w
a_1	λ_{11}	λ_{12}	$\cdots$	λ_{1j}	$\cdots$	λ_{1w}
a_2	λ_{21}	λ_{22}	$\cdots$	λ_{2j}	$\cdots$	λ_{2w}
$\cdots$						
a_i	λ_{i1}	λ_{i2}	$\cdots$	λ_{ij}	$\cdots$	λ_{iw}
$\cdots$						
a_t	λ_{t1}	λ_{t2}	$\cdots$	λ_{tj}	$\cdots$	λ_{tw}

TABLE II
THE LOSS FUNCTION OF THE PARKING PROBLEMS

Actions \ States	$s_1(\leq 2hours)$	$s_2(> 2hours)$
$a_1(park\ on\ meter)$	\$2	\$12
$a_2(park\ in\ a\ parking\ lot)$	\$7	\$7

Given an object x, let $\boldsymbol{x}$ is its description. $\boldsymbol{\Omega} = \{s_1, s_2, \cdots, s_w\}$ is a finite set of w states that x is possibly in, and $\boldsymbol{A} = \{a_1, a_2, \cdots, a_t\}$ is a finite collection of t possible actions. Let $P(s_j|\boldsymbol{x})$ is the conditional probability of x being in state s_j, and the loss function $\lambda(a_i|s_j)$ denotes the loss (or cost) for taking the action a_i when the state is s_j.

All the values of loss functions can be illustrated in Table I, with the columns denoting the set $\boldsymbol{A}$ of t actions and the rows denoting the set $\boldsymbol{\Omega}$ of w states. $\lambda(a_i|s_j)$ denotes the loss (or cost) for taking the action a_i when the state is s_j. It can be simplified as λ_{ij} .

For an object x with description $\boldsymbol{x}$, supposing action a_i is taken. The expected loss associated with action a_i is calculated by the following equation

$$R(a_i|\boldsymbol{x}) = \sum_{j=1}^{w} \lambda_{ij} P(s_j|\boldsymbol{x}) \tag{1}$$

$R(a_i|\boldsymbol{x})$ is called the conditional risk. Given the loss functions and the probabilities, one can compute the expected loss of a certain action. Furthermore, comparing the expected loss of all the actions, one can decide a specified action with the minimum loss.

Example: The procedure of Bayesian decision theory can be illustrated by the following example [31]. Suppose, there are two states: s_1 indicates that a meeting will be not more than 2 hours, and s_2 indicates that the meeting will be more than 2 hours. That is, $\boldsymbol{\Omega} = \{s_1, s_2\}$. Obviously, the two states are complement. Suppose the probability of appearance of s_1 is 0.8, then the probability of appearance of s_2 is 0.2, namely

$$P(s_1|\boldsymbol{x}) = 0.80 \tag{2}$$

$$P(s_2|\boldsymbol{x}) = 1 - P(s_1|\boldsymbol{x}) = 0.20 \tag{3}$$

Let $\boldsymbol{A} = \{a_1, a_2\}$ is the collection of actions, where a_1 represent the car drove by a participant will be parked on meter, and a_2 represent the car will be parked in the parking lot. The loss functions for taking different actions in different states can be expressed as Table II.

In this case, the expected loss $R(a_i|\boldsymbol{x})$ associated with

taking the individual action can be expressed as

$$\begin{aligned} R(a_1|\boldsymbol{x}) &= \sum_{j=1}^{2}(\lambda_{1j}P(s_j|\boldsymbol{x})) \\ &= \lambda_{11}P(s_1|\boldsymbol{x})) + \lambda_{12}P(s_2|\boldsymbol{x})) \\ &= \$2 * 0.80 + \$12 * 0.20 \\ &= \$3 \end{aligned} \tag{4}$$

Similarly,

$$\begin{aligned} R(a_2|\boldsymbol{x}) &= \$7 * 0.80 + \$7 * 0.20 \\ &= \$7 \end{aligned} \tag{5}$$

According to $\$3 < \$7$, it is reasonable that the car will be parked on meter by the participant.

The process of inference based on Bayesian decision theory, which not only considers the loss of misjudgment, but it also considers the probability of appearance of various states. Therefore, it can make the preferable decision according to the actual situation.

III. The proposed methodology

In this section, Bayesian decision theory will be applied to the field of clustering and the risk assessment function of clustering scheme will be constructed firstly. Then, the method to select the initial seeds for K-means based on the max-min distance will be presented. According to the mentioned content, the adaptive clustering algorithm will be designed in the end.

A. Evaluating the risk of a clustering scheme

In order to construct the risk assessment function of a clustering scheme, Bayesian decision theory will be applied to the field of clustering firstly.

1) Extending Bayesian decision theory to the field of clustering: We will evaluate the risk of a clustering scheme through evaluating the risk of the state of clustering two objects. That is, the object mentioned in Section II-B is a single object x while the object investigated here is a pair (x_i, x_j).

Let $\boldsymbol{\Omega} = \{C_1, C_2\}$ denote the set of states indicating that the two objects x_i and x_j are in the same cluster and in the different cluster, respectively.

Let $\boldsymbol{A} = \{a_1, a_2\}$ be the collection of two possible actions, where a_1 denotes the action to clustering objects x_i and x_j into the same cluster, and a_2 denotes the action to clustering objects x_i and x_j not into the same cluster. That is, a_1 and a_2 denote the action to clustering a pair (x_i, x_j) into the state C_1 and C_2, respectively.

Let λ_{ij} denote the loss for taking the action a_i when the state is C_j. where $i = 1, 2;\ j = 1, 2$.

For a pair (x_i, x_j) with the description $(\boldsymbol{x}_i, \boldsymbol{x}_j)$, when clustering x_i and x_j into the same cluster or not into the same cluster, the expected risks can be respectively expressed as follows:

$$\begin{aligned} R(a_1|(\boldsymbol{x}_i, \boldsymbol{x}_j)) &= \lambda_{11}P(C_1|(\boldsymbol{x}_i, \boldsymbol{x}_j)) + \lambda_{12}P(C_2|(\boldsymbol{x}_i, \boldsymbol{x}_j)) \\ R(a_2|(\boldsymbol{x}_i, \boldsymbol{x}_j)) &= \lambda_{21}P(C_1|(\boldsymbol{x}_i, \boldsymbol{x}_j)) + \lambda_{22}P(C_2|(\boldsymbol{x}_i, \boldsymbol{x}_j)) \end{aligned} \tag{6}$$

Without loss of generality, we can consider the range of loss function be $[0, 1]$. Obviously, the case of considering two endpoints of the interval $[0, 1]$ is able to distinguish losses more clearly, thus, we can set $\lambda_{11} = 0$, $\lambda_{12} = 1$, $\lambda_{21} = 1$, $\lambda_{22} = 0$. That is, clustering objects belonging to a cluster to the same cluster without loss, namely 0; clustering objects belonging to different clusters to the same cluster with maximum loss, namely 1. Similarly, the loss of clustering objects belonging to same cluster to the different clusters is also maximum, namely 1; the loss of clustering objects belonging to different clusters to the different clusters is minimal, namely, 0. Thus, Eq.(6) can be simplified as Eq.(7).

$$\begin{aligned} R(a_1|(\boldsymbol{x}_i, \boldsymbol{x}_j)) &= P(C_2|(\boldsymbol{x}_i, \boldsymbol{x}_j)) \\ R(a_2|(\boldsymbol{x}_i, \boldsymbol{x}_j)) &= P(C_1|(\boldsymbol{x}_i, \boldsymbol{x}_j)) \end{aligned} \tag{7}$$

In a clustering scheme CS, let $R(CS|(\boldsymbol{x}_i, \boldsymbol{x}_j))$ denote the risk to clustering (x_i, x_j). If the objects x_i and x_j are grouped into the same cluster in CS, then

$$R(CS|(\boldsymbol{x}_i, \boldsymbol{x}_j)) = P(C_2|(\boldsymbol{x}_i, \boldsymbol{x}_j)) \tag{8}$$

otherwise

$$R(CS|(\boldsymbol{x}_i, \boldsymbol{x}_j)) = P(C_1|(\boldsymbol{x}_i, \boldsymbol{x}_j)) \tag{9}$$

2) Constructing the risk assessment function of a clustering scheme: In a clustering problem, we usually consider a similarity matrix S. Let $s(x_i, x_j)$ denote the similarity of two objects x_i and x_j. In the paper, $s(x_i, x_j)$ can be calculated by the following formula

$$s(x_i, x_j) = 1 - \frac{d(x_i, x_j)}{\max\limits_{i,j} d(x_i, x_j)} \tag{10}$$

where $d(x_i, x_j)$ denote the Euclidean distance between x_i and x_j, and $\max\limits_{i,j} d(x_i, x_j)$ denote the maximum value of the Euclidean distance for all pairs of objects.

In the similarity matrix S, there must exist a threshold which can measure the two objects are similar or not. Let v denotes the value of the threshold. In general, there are many ways to get the value of v. Such as to get the value by human defining, to set the value based on the physical sense in the data, to get the value by other statistic methods and so on. In the paper, we use the statistical methods to set the average of similarity among all objects as the value of v , namely

$$v = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} s(x_i, x_j) \tag{11}$$

In the similarity matrix S, if $s(x_i, x_j) \geq v$, then we can set that $P(C_1|(\boldsymbol{x}_i, \boldsymbol{x}_j)) \geq 0.5$. According to the similarity $s(x_i, x_j)$ is larger, the two objects are clustered into the same cluster is more possible, it is reasonable to suppose that the $P(C_1|(\boldsymbol{x}_i, \boldsymbol{x}_j))$ is proportional to the $s(x_i, x_j)$. Thus, the following equation can be constructed to calculate the probability

$$P(C_1|(\boldsymbol{x}_i, \boldsymbol{x}_j)) = \begin{cases} 0.5 + \frac{s(x_i,x_j)-v}{2(1-v)} & s(x_i, x_j) \geq v \\ 0.5 - \frac{v-s(x_i,x_j)}{2v} & s(x_i, x_j) < v \end{cases} \tag{12}$$

Correspondingly, we can get $P(C_2|(\boldsymbol{x}_i, \boldsymbol{x}_j))$ as follow

$$\begin{aligned} P(C_2|(\boldsymbol{x}_i, \boldsymbol{x}_j)) &= 1 - P(C_1|(\boldsymbol{x}_i, \boldsymbol{x}_j)) \\ &= \begin{cases} 0.5 - \frac{s(x_i,x_j)-v}{2(1-v)} & s(x_i, x_j) \geq v \\ 0.5 + \frac{v-s(x_i,x_j)}{2v} & s(x_i, x_j) < v \end{cases} \end{aligned} \tag{13}$$

$$R(CS|(\boldsymbol{x}_i,\boldsymbol{x}_j)) = \begin{cases} 0.5 - \frac{s(x_i,x_j)-v}{2-2v} & s(x_i,x_j) \geq v, x_i \text{ and } x_j \text{ are in the same cluster in CS} \\ 0.5 + \frac{v-s(x_i,x_j)}{2v} & s(x_i,x_j) < v, x_i \text{ and } x_j \text{ are in the same cluster in CS} \\ 0.5 + \frac{s(x_i,x_j)-v}{2-2v} & s(x_i,x_j) \geq v, x_i \text{ and } x_j \text{ are in different clusters in CS} \\ 0.5 - \frac{v-s(x_i,x_j)}{2v} & s(x_i,x_j) < v, x_i \text{ and } x_j \text{ are in different clusters in CS} \end{cases} \tag{14}$$

Substitute Eq.(12) to Eq.(9) or Substitute Eq.(13) to Eq.(8), we can get the specific risk $R(CS|(\boldsymbol{x}_i,\boldsymbol{x}_j))$ as shown in Eq.(14).

Considering all pairs in a clustering scheme CS , the risk of the clustering scheme CS can be evaluated as follow:

$$R(CS) = \sum_{i=1}^{n}\sum_{j=1}^{n} R(CS|(\boldsymbol{x}_i,\boldsymbol{x}_j)) \tag{15}$$

From Eq.(14), we can know that if $s(x_i,x_j) \geq v$, the risk of clustering (x_i,x_j) into the same cluster does not exceed 0.5 while the risk of clustering them into different clusters is more than 0.5. Similarly, if $s(x_i,x_j) < v$, the risk of clustering (x_i,x_j) into different clusters does not exceed 0.5 while the risk of clustering them into the same cluster is more than 0.5. Namely, the risk of clustering the objects which are not similar to each other into the same cluster is larger than the risk of clustering them into different clusters; the risk of clustering the objects which are similar to each other into different clusters is larger than the risk of clustering them into the same clusters.

The goal of clustering is to group data points into clusters such that the data in each cluster shares a high degree of similarity while being very dissimilar to data from other clusters. According to the analysis mentioned in the previous paragraph, We can know that the closer the clustering scheme is to the real classification, the smaller the value of the risk assessment function. In other words, the value of the risk assessment function is smaller, the quality of the clustering scheme is better.

B. A method to select initial seeds for K-means based on max-min distance

The selection of initial seeds of K-means clustering algorithm is quite stochastic, which leads to the fact that the outcome of clustering is also quite stochastic. In order to obtain a certain clustering result, we propose a new method to select initial seeds for K-means based on max-min distance.

the basic idea of the method is shown as follow. If there is only one cluster, that is, all data objects are in the same cluster. The object which has the minimum contribution to total intra-cluster distance is most likely to become the cluster center. Thus, we can compute the contribution of each data object to total intra-cluster distance, then select the object with the minimum value as the first initial seed c_1. If all data objects are divided to two clusters, we can choose the object which are far away from c_1 as the second initial seed c_2. Because these centers selected in this way are not too concentrated, it is helpful that to obtain a better result of clustering. Similarly, if data set is grouped into three clusters, we can choose the object which is far away from c_1 and c_2 as the next initial seed. Do this until all the required initial seeds are found. Now we describe the details of the method as Algorithm 2.

Algorithm 2 Selection of initial centers

Input: data set $X = \{x_1, x_2, \cdots, x_n\}$, the number of clusters K ($K \geq 2$).

Output: the set of initialization centers C.

1 Compute the contribution of each data point to total intra-cluster distance and then select the object with the smallest one as the first initial center c_1. Namely, $C = \{c_1\}$
2 Select the object that is furthest away from c_1 as the second initial center c_2. $C = C \cup \{c_2\}$
3 **if** $K = 2$, **then**
4 output $C = \{c_1, c_2\}$, end the algorithm.
5 **else**
6 **for** $i = 3$ to K **do**
7 set $c_i = x_k$ such that $d(c_i, c_j) = \max_{x_k}(\min_{c_j}(d(x_k, c_j))$, where $x_k \in X - C, c_j \in C, j = 1, 2, 3, \cdots, i-1$.
8 $C = C \cup \{c_i\}$
9 **end for**
10 **end if**

C. The adaptive clustering algorithm

In order to improve the quality of the result of clustering, the original data is processed firstly. Subsequently, we propose the adaptive clustering algorithm based on Bayesian decision theory and max-min distance in this subsection.

1) Processing the original data: Because clustering does not require any information apart from the given data set, we usually do not know which attributes are important and which attributes are not important. Thus, we present that the value of attribute belonging to the same attribute should be normalized within the range of $0-1$. The goal of normalization is to give the same emphasize on each property regardless of their actual domain sizes.

For the jth attribute value of the ith data object x_{ij}, the normalized value $x_{ij}^{'}$ is computed as Eq.(16)

$$x_{ij}^{'} = \frac{x_{ij} - n}{m - n + \varepsilon} \tag{16}$$

where m and n denote the maximum and minimum domain values of the jth attribute, respectively. ε is a very small number, and it added to denominator is to ensure that the fraction is meaningful. In the paper, ε is set as 0.0000001.

2) The adaptive clustering algorithm: Based on the analysis of the above content, we can design an adaptive clustering algorithm. The new algorithm adopts the idea of continuous partition of a given data set. In the process of each partition, the algorithm can select initial seeds based on max-min distance to obtain a certain clustering scheme, and it can evaluate the risk of the clustering scheme by the proposed risk assessment function. Comparing the risk values before and after partitioning, the algorithm can decide whether the

data set is continue partitioned, thus it can determine the number of clusters and get the final result of clustering automatically. The specific steps of the proposed algorithm are shown in Algorithm 3.

Algorithm 3 Adaptive clustering algorithm

Input: Data set $X = \{x_1, x_2, \cdots, x_n\}$.
Output: The result of clustering CS.

1 The attribute value of input data objects is normalized by Eq.(16).
2 Computing similarity matrix S using Eq.(10).
3 Setting all of the objects in the same cluster, namely $K = 1$, $CS_1 = \{C_1 = \{x_1, x_2, \cdots, x_n\}\}$.
4 If all the elements in S are equal, then output the CS_1, end the algorithm; Otherwise, computing the value of v according to Eq.(11). Computing the matrix P according to Eq.(12), and computing the risk $R(CS_1)$ according to Eq.(15), set $R(CS_u) = R(CS_1)$, go to step 5.
5 Set $K = K + 1$, determining the initial centers C using Algorithm 2, and performing K-means clustering algorithm based on C to obtain the clustering scheme CS_w, then computing the risk $R(CS_w)$, go to step 6.
6 If $R(CS_w) \leq R(CS_u)$, set $R(CS_u) = R(CS_w)$, go to step 5; Otherwise, output the CS_u, end the algorithm.

IV. Experimentation

In this section, some evaluation methods for clustering result are introduced as well as experimental results and the corresponding analysis are illustrated.

A. Evaluation methods for clustering result

To evaluate the efficiency of clustering algorithms, three evaluation indexes— purity, rand index, and the number of clusters are employed in the following experiments. The purity [32] and the rand index [33] are respectively calculated as follows:

•$Purity$

The purity measure is an external evaluation criterion that evaluates the quality of the clusters according to the labeled samples available. In clustering, a cluster is considered as pure if it contains labeled data points from one and only one class. On the contrary, a cluster is considered as impure if it contains labeled data points from many different classes. The purity is computed by the following formula:

$$Purity(S, CS) = \frac{1}{|U|} \sum_i \max_j |C_i \cap W_j| \qquad (17)$$

where $S = \{W_1, W_2, \ldots, W_s\}$ is the set of the true classes, $CS = \{C_1, C_2, \ldots, C_r\}$ is the set of clusters, and $|U|$ denotes that the size of the data set. Bad clustering schemes have purity values close to 0 while a perfect clustering scheme has a purity of 1.

•$Rand\,Index$

Given a set of n elements $U = \{x_1, \cdots, x_n\}$ and two partitions of U to compare, $C = \{C_1, \ldots, C_r\}$, a partition of U into r subsets, and $S = \{S_1, \ldots, S_s\}$, a partition of U into s subsets. In order to define the rand index, the following quantities are needed:

a is the set of pairs of points that belong to the same class and that are clustered in the same cluster;
b is the set of pairs of points that belong to different classes and that are clustered in different clusters;
c is the set of pairs of points that belong to the same class but are placed in different clusters;
d is the set of pairs of points that belong to different classes but are placed in the same cluster;

The rand index is computed as

$$Rand\,Index = \frac{a+b}{a+b+c+d} \qquad (18)$$

Intuitively, $a + b$ can be considered as the number of agreements between C and S, and $c + d$ as the number of disagreements between C and S.

In the paper, C denotes the clustering scheme obtained by different clustering method that employed in the following experiments, and S presents the true classification of real world data sets. The greater the value of rand index, the clustering scheme the more similar to the true classification.

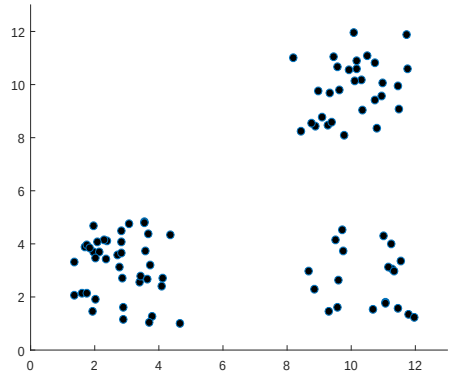

Fig. 1. The two-dimensional data set X90

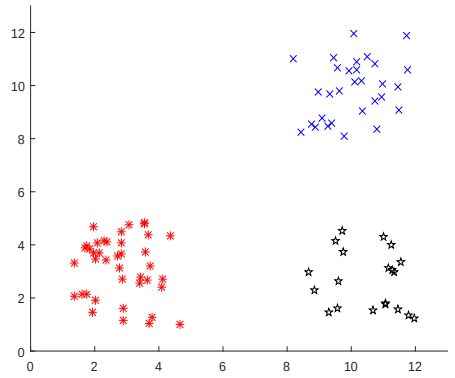

Fig. 2. The clustering result of proposed method on the data set X90

B. Evaluation on clustering effectiveness

In this subsection, we have carried out many experiments on several synthetic and real world data sets to highlight the performance of the proposed algorithm.

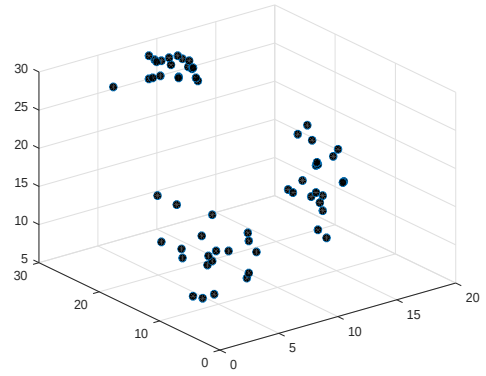

Fig. 3. The three-dimensional data set X60

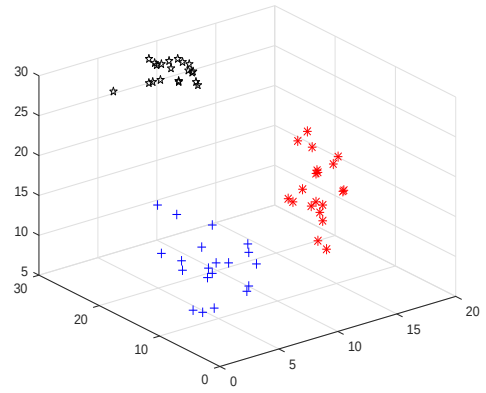

Fig. 4. The clustering result of proposed method on the data set X60

1) Visualize the clustering results: In order to show that the proposed clustering algorithm can be put into operation effectively and can get a reasonable clustering result, two artificial data sets are used to visualize the results of clustering, one is two-dimensional and the other is three-dimensional.

The two-dimensional data set called X90 is shown in Fig.1. It contains 90 objects grouped into 3 distinct classes. The number of elements in the 3 categories are 40, 30 and 20, respectively. The clustering result of the proposed method is shown in Fig.2. It illustrates that the number of clusters of the data set X90 determined by the proposed method automatically is 3, which coincides with the fact that there are 3 classes in the raw data set. Comparing Fig.1 and Fig.2, we can know that the proposed method can obtain reasonable clustering result.

Fig.3 describes an artificial three-dimensional data set, named X60, which contains 60 objects assigned to 3 classes. In the 3 classes, the number of elements in each class is 20. The experimental result of the proposed method is shown in Fig.4. It also shows that the proposed method can be terminated at the right number of clusters automatically and get good clustering result.

Although the data sets X90 and X60 are very simple, they are able to illustrate the effectiveness of the proposed algorithm in a certain extent. K-means with K=3 may easily cluster them perfectly. However, the value of K needs to specified by human in advance while the proposed method can determine it automatically. The experiments on synthetic data sets show that the proposed method, without human interferences, can determine the number of clusters and get satisfactory clustering result.

TABLE III
A BRIEF DESCRIPTION OF DATA SETS

Datasets	Instances	Attributes	Classes
Iris	150	4	3
Seeds	210	7	3
E.coli	336	8	8
Image Segment	2310	19	7
Page Blocks	5473	10	5
Landsat	2000	37	6

TABLE IV
THE COMPARISONS OF THE PURITY OF CLUSTERING RESULTS

Model	Ours(%)	K-means(%)	K-means++(%)
Iris	**88.6667**	84.2207	84.7184
Seeds	88.0952	88.8519	**89.0467**
E.coli	**77.6786**	56.1554	59.0899
Image Segment	57.9221	59.3221	**59.6178**
Page Blocks	**89.5304**	48.9566	46.9419
Landsat	**73.5000**	66.7352	66.7921

2) Comparative experiments: In order to verify the performance of the proposed method, six data sets are downloaded from UCI machine learning data repository [34]. A brief description of these data sets is shown in Table III. More detailed description of these data sets can be found in [34].

Some experiments are done on the data sets mentioned above, and the experimental results are analyzed by the evaluation methods mentioned in Section 4.1.

Due to K-means clustering algorithm is adopted in the framework of the proposed method, thus, we compare the proposed method with the K-means clustering algorithm and K-means++ clustering algorithm that is an improved K-means clustering algorithm and is used widely in partition clustering algorithm. The comparison of the proposed method with K-means algorithm and K-means++ algorithm using purity and rand index are shown in Table IV and Table V, respectively. Because K-means algorithm and K-means++ algorithm require to specify the number of clusters, we set the true number of classes in each given data set as the value of K. In addition, as the initial seeds of K-means and the first initial seed of K-means++ are selected quite stochastic, which may lead to the fact that the result of clustering are also stochastic, we carry out 1000 runs of the K-means algorithm and K-means++ algorithm on these standard data sets respectively and take the average of the evaluation indexes of 1000 times experiments as final

TABLE V
THE COMPARISONS OF THE RAND INDEX OF CLUSTERING RESULTS

Model	Ours(%)	K-means(%)	K-means + +(%)
Iris	**87.3736**	85.1180	85.4331
Seeds	84.0829	86.7166	**86.9264**
E.coli	**87.3810**	80.4468	81.2503
Image Segment	82.1867	85.8635	**85.9533**
Page Blocks	**48.2843**	45.5189	44.5283
Landsat	**85.8425**	85.2427	85.2970

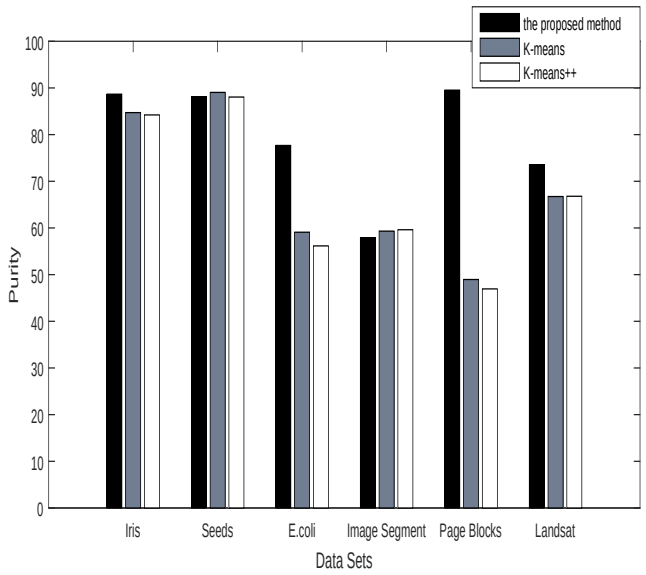

Fig. 5. Comparison of the purity of three algorithms on six data sets

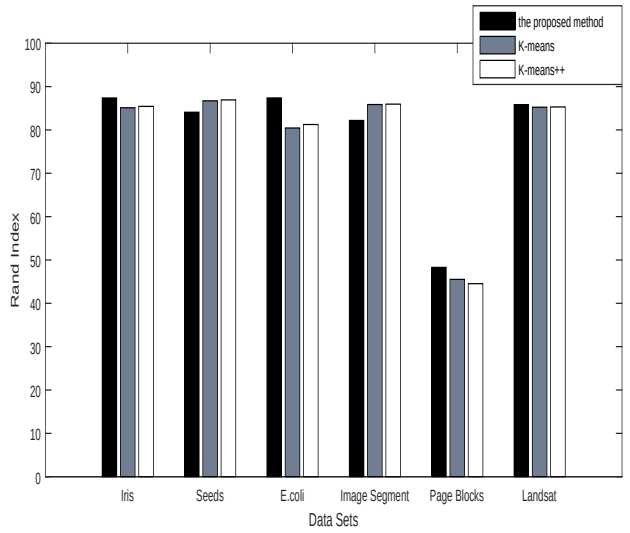

Fig. 6. Comparison of the rand index of three algorithms on six data sets

evaluation indexes.

In order to compare the experimental results more intuitively, we use the data in Table IV and Table V to generate the bar graph as shown in Fig. 5 and Fig. 6, respectively. Observing Table IV, Table V, Fig. 5 and Fig. 6,we can find that the quality of clustering results of the proposed method outperforms the results of the other two methods on the most data sets. Which shows that the proposed method is able to obtain efficient clustering results.

Since the proposed method can automatically determine the number of clusters while K-means and K-means++ require human to specify the value of K, other two methods which are usually used to determine the number of clusters are compared with the proposed method. One is best K-means algorithm, which determines the number using elbow method based on the framework of K-means clustering algorithm. The other is FACA-DTRS algorithm, which determines the cluster number using decision-theoretic rough set based on the framework of hierarchical agglomerative clustering algorithm.

TABLE VI
THE COMPARISONS OF THE NUMBER OF CLUSTERS

Datasets	Ours	FACA-DTRS	Best-kmeans	TrueCluster
Iris	3	3	4	3
Seeds	4	3	8	3
E.coli	7	6	35	8
Image Segment	5	80	21	7
Page Blocks	6	*	*	5
Landsat	7	14	27	6

Table VI records the comparison of the number of clusters obtained by our proposed method and the two methods. Because the time complexity of the two methods are too high to get the cluster numbers of Page Blocks data set, we record the results as " $*$ ". To observe Table VI, the number of clusters from our proposed method is much closer to the right number of classes than the other two methods. Which illustrates that the proposed method can determine the reasonable number of clusters adaptively.

V. CONCLUSION

The goal of this paper is to development an approach for adaptive clustering. In order to achieve this goal, Bayesian decision theory is applied to the field of clustering firstly. Then a risk assessment function is constructed to evaluate the risk of clustering scheme. Furthermore, in order to obtain a certain results of clustering, we presented a method to select initial seeds based on max-min distance. Finally, we designed an adaptive clustering algorithm based on the above mentioned content. To study the effectiveness of the proposed method for clustering, we conducted extensive experiments. The experimental results show that the proposed method is able to produce more efficient clustering results than traditional K-means and K-means++.

For future work, we attempt to devise a method to obtain better initial seeds so that the idea of the paper can be applied to more complicated data sets and get better clustering results.

ACKNOWLEDGMENTS

The authors thank the editors and the anonymous reviewers for helpful comments and suggestions.

REFERENCES

[1] J. Han and M. Kamber, "Data mining: Concepts and techniques, morgan kaufmann," *Machine Press, 2001 (in Chinese)*, vol. 5, no. 4, pp. 394–395, 2006.

[2] Z. Liu, Q. Zheng, L. Xue, and X. Guan, "A distributed energy-efficient clustering algorithm with improved coverage in wireless sensor networks ," *Future Generation Computer Systems*, vol. 28, no. 5, pp. 780–790, 2012.

[3] Y. Guo and A. Sengur, "Ncm: Neutrosophic c-means clustering algorithm," *Pattern Recognition*, vol. 48, no. 8, pp. 2710–2724, 2015.

[4] I. C. M. Mario, R. P. Pablo, and R. A. Graciela, "A fuzzy clustering approach for face recognition based on face feature lines and eigenvectors," Engineering Letters, vol. 15, no. 1, pp 35–44, 2007.

[5] X. Zhong, G. Yang, L. Li, and L. Zhong, "Clustering and correlation based collaborative filtering algorithm for cloud platform," IAENG International Journal of Computer Science, vol. 43, no. 1, pp 108–114, 2016.

[6] T. Sakai, K. Tamura, and H. Kitakami, "Extracting attractive local-area topics in georeferenced documents using a new density-based spatial clustering algorithm," IAENG International Journal of Computer Science, vol. 41, no. 3, pp 185–192, 2014.

[7] L. N. Chi and H. S. Nguyen, "A method of web search result clustering based on rough sets," in *Ieee/wic/acm International Conference on Web Intelligence, 2005. Proceedings*, 2005, pp. 673–679.
[8] D. Jiang, C. Tang, and A. Zhang, "Cluster analysis for gene expression data: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 11, pp. 1370–1386, 2004.
[9] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*. Prentice Hall, 1988.
[10] J. Senthilnath, D. Kumar, J. A. Benediktsson, and X. Zhang, "A novel hierarchical clustering technique based on splitting and merging," *International Journal of Image and Data Fusion*, vol. 7, pp. 1–23, 2015.
[11] D. W. Choi and C. W. Chung, "A k-partitioning algorithm for clustering-scale spatio-textual data," *Information Systems*, vol. 64, pp. 1–11, 2016.
[12] L. Duan, L. Xu, F. Guo, J. Lee, and B. Yan, "A local-density based spatial clustering algorithm with noise," in *IEEE International Conference on Systems, Man and Cybernetics*, 2006, pp. 4061–4066.
[13] J. Galvao, M. Y. Santos, J. M. Pires, and C. Costa, "Dealing with repeated objects in snnagg," IAENG International Journal of Computer Science, vol. 43, no. 1, pp 115–125, 2016.
[14] I. Foster, T. Freeman, K. Keahy, D. Scheftner, B. Sotomayer, and X. Zhang, "Virtual clusters for grid communities," in *IEEE International Symposium on CLUSTER Computing and the Grid*, 2006, pp. 513–520.
[15] M. A. Murphy and S. Goasguen, "Virtual organization clusters: Self-provisioned clouds on the grid," *Future Generation Computer Systems*, vol. 26, no. 8, pp. 1271–1281, 2010.
[16] J. B. Macqueen, "Some methods for classification and analysis of multivariate observations," in *Proc. of Berkeley Symposium on Mathematical Statistics and Probability*, 1967, pp. 281–297.
[17] S. P. Lloyd, "Least squares quantization in pcm," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.
[18] S. Z. Selim and M. A. Ismail, "K-means-type algorithms: A generalized convergence theorem and characterization of local optimality," *IEEE Trans Pattern Anal Mach Intell*, vol. 6, no. 1, pp. 81–87, 1984.
[19] M. B. Al-Daoud and S. A. Roberts, "New methods for the initialisation of clusters," *Pattern Recognition Letters*, vol. 17, no. 5, pp. 451–455, 1996.
[20] D. Arthur and S. Vassilvitskii, "k-means++:the advantages of careful seeding," in *Eighteenth Acm-Siam Symposium on Discrete Algorithms, SODA 2007, New Orleans, Louisiana, Usa, January*, 2015, pp. 1027–1035.
[21] I. Katsavounidis, C. C. Jay Kuo, and Z. Zhang, "A new initialization technique for generalized lloyd iteration," *Signal Processing Letters IEEE*, vol. 1, no. 10, pp. 144–146, 1994.
[22] S. S. Khan and A. Ahmad, "Cluster center initialization algorithm for k-means clustering," *Expert Systems with Applications*, vol. 25, no. 11, pp. 1293–1302, 2014.
[23] S. J. Redmond and C. Heneghan, "A method for initialising the k-means clustering algorithm using kd-trees," *Pattern Recognition Letters*, vol. 28, no. 8, pp. 965–973, 2007.
[24] J. Dong and M. Qi, "K-means optimization algorithm for solving clustering problem," in *Second International Workshop on Knowledge Discovery and Data Mining*, 2009, pp. 52–55.
[25] S. Gajawada and D. Toshniwal, "Projected clustering using particle swarm optimization," *Procedia Technology*, vol. 4, pp. 360–364, 2012.
[26] G. P. Babu and M. N. Murty, "A near-optimal initial seed value selection in k-means algorithm using a genetic algorithm," *Pattern Recognition*, vol. 14, no. 10, pp. 763–769, 1993.
[27] M. Laszlo and S. Mukherjee, "A genetic algorithm that exchanges neighboring centers for k-means clustering," *Pattern Recognition Letters*, vol. 28, no. 16, pp. 2359–2366, 2007.
[28] H. Yu, Z. Liu, and G. Wang, "An automatic method to determine the number of clusters using decision-theoretic rough set," *International Journal of Approximate Reasoning*, vol. 55, no. 1, pp. 101–115, 2014.
[29] "About matlab [online]," Available:http://www.mathworks.com/.
[30] M. R. Rezaee, B. P. F. Lelieveldt, and J. H. C. Reiber, "A new cluster validity index for the fuzzy c-mean," *Pattern Recognition Letters*, vol. 19, no. 3-4, pp. 237–246, 1998.
[31] Y. Yao and Y. Zhao, "Attribute reduction in decision-theoretic rough set models," *Information Sciences*, vol. 178, no. 17, pp. 3356–3373, 2008.
[32] C. D. Manning, P. Raghavan, and H. Schtze, "Introduction to information retrieval: Hierarchical clustering," 2008.
[33] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, 1971.
[34] "Ucivine machine learning repository," http://archive.ics.uci.edu/ml/.

Fengqin Zhao was born in Tai'an, Shandong Province, China in 1991. She received her B. S. degree in the Department of Mathematics and Applied Mathematics from Taishan University in 2014, and she obtained her M. S. degree in the Department of Mathematics from Xidian University in 2017. She is major in Probabilistic graphical models, data analysis and its application.

Youlong Yang received his B. S. and M. S. in Mathematics from Shaanxi Normal University, Xian, China in 1990 and 1993 respectively, and Ph.D. in System Engineering from Northwester Polytechnical University, Xian,China in 2003. Since 2004, he has been with the faculty at Xidian University, Xi'an, China. His research interests include Machine learning, Statistical data analysis and Probabilistic graphical models.

Weiwei Zhao was born in Zhangjiakou, Hebei Province, China in 1989. She received her B. S. degree in the Department of Mathematics and Applied Mathematics from Langfang Teachers College in 2014. And began work for a master at Xidian University in 2014. She is interested in Data Analysis and Machine Learning.