Computers and Mathematics with Applications 58 (2009) 474–483

Contents lists available at ScienceDirect

Computers and Mathematics with Applications

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

An initialization method for the *K*-Means algorithm using neighborhood model

Fuyuan Cao [a,b], Jiye Liang [a,b,*], Guang Jiang [c]

[a] School of Computer and Information Technology, Shanxi University, Taiyuan, 030006, Shanxi, China

[b] Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Taiyuan, 030006, China

[c] Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China

ARTICLE INFO

Article history:
Received 6 January 2008
Received in revised form 23 March 2009
Accepted 8 April 2009

Keywords:
Neighborhood model
Initial cluster centers
Cohesion degree
Coupling degree
K-Means clustering algorithm

ABSTRACT

As a simple clustering method, the traditional *K*-Means algorithm has been widely discussed and applied in pattern recognition and machine learning. However, the *K*-Means algorithm could not guarantee unique clustering result because initial cluster centers are chosen randomly. In this paper, the cohesion degree of the neighborhood of an object and the coupling degree between neighborhoods of objects are defined based on the neighborhood-based rough set model. Furthermore, a new initialization method is proposed, and the corresponding time complexity is analyzed as well. We study the influence of the three norms on clustering, and compare the clustering results of the *K*-means with the three different initialization methods. The experimental results illustrate the effectiveness of the proposed method.

1. Introduction

Clustering is one of the widely used knowledge discovery techniques to reveal structures in a data set that can be extremely useful to the analyst [1]. As clustering do not make any statistical assumptions to data, it is referred to as unsupervised learning algorithm. In general, the problem of clustering deals with partitioning a data set consisting of n points embedded in m-dimensional space into k distinct set of clusters, such that the data points within the same cluster are more similar to each other than to data points in other clusters. A common method is to use data to learn a set of centers such that the sum of squared errors between objects and their nearest centers is small. Clustering techniques are generally classified as partitional clustering and hierarchical clustering, based on the properties of the generated clusters. The partitional clustering technique usually begins with an initial set of randomly selected exemplars and iteratively refines this set so as to decrease the sum of squared errors. Due to the simpleness, random initialization method has been widely used. However, partitional clustering algorithms with random initialization method need to be rerun many times with different initializations in an attempt to find a good solution. Furthermore, random initialization method works well only when the number of clusters is small and chances are good that at least one random initialization is close to a good solution [2]. Therefore, how to choose proper initial cluster centers is extremely important as they have a direct impact on the formation of final clusters.

The *K*-Means clustering algorithm [3], developed three decades ago, is one of the best-known and most popular clustering algorithms used in a variety of domains. Despite being used in a wide array of applications, the *K*-Means algorithm is not exempt from drawbacks. Some of these drawbacks have been extensively reported in some literatures. The most important are listed below [4]:

* Corresponding author at: School of Computer and Information Technology, Shanxi University, Taiyuan, 030006, Shanxi, China.
E-mail addresses: cfy@sxu.edu.cn (F. Cao), ljy@sxu.edu.cn (J. Liang), jiangg_211@126.com (G. Jiang).

0898-1221/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.camwa.2009.04.017

1. As many clustering methods, the K-Means algorithm assumes that the number of clusters K is already known by the users, which, unfortunately, usually is not true in practice.
2. As an iterative technique, the K-Means algorithm is especially sensitive to initial cluster centers.
3. The K-Means algorithm converges to a local minima.

Although there is no guarantee of achieving a global minima, at least the convergence of the K-Means algorithm is ensured [5]. Therefore, how to choose proper initial cluster centers becomes very important for the K-Means algorithm. The problem of initial cluster centers is not exclusive to the K-Means algorithm but shared with many clustering algorithms that work as a hill-climbing strategy whose deterministic behavior leads to a local minima dependent on initial cluster centers. Several attempts have been made to solve the cluster initialization problem. A recursive method for initializing the means by running K clustering problems is discussed by Duda and Hart [6]. Milligan [7] showed the strong dependence of the K-Means algorithm on initial clustering and suggested that good final cluster structures can be obtained using Ward's hierarchical method [8] to provide the K-Means algorithm with initial clusters. Fisher [9] proposed creating the initial cluster centers by constructing an initial hierarchical clustering based upon the work [10]. Higgs et al. [11] and Snarey et al. [12] suggested using a MaxMin algorithm in order to select a subset of the original database as the initial centroid to establish the initial clusters. Bradley et al. [13] used bilinear program to determine k initial clusters such that the sum of distances of each point to the nearest center is minimized. Bradley and Fayyad [14] presented a procedure for computing a refined starting condition from a given initial one that is based on an efficient technique for estimating the modes of a distribution. Khan and Ahmad [15] presented an algorithm for computing initial cluster centers for K-Means algorithm. Pená et al. [4] presented a comparative study for different initialization methods for the K-Means algorithm, the result of their experiments illustrate that the random and Kaufman initialization method outperforms the rest of the compared methods as they make the K-Means algorithm more effective and more independent on initial clustering and on instance order. However, there are no universally accepted method for selecting initial cluster centers as reported by Meila and Heckerman [16].

Rough set theory introduced by Pawlak [17] is a kind of symbolic machine learning technology for categorical value information systems with uncertainty information [18,19]. In recent years, rough set theory has attracted much attention in clustering literatures. Parmar et al. [20] proposed a new algorithm MMR(Min-Min-Roughness) for clustering categorical data based on rough set theory, which has the ability to handle the uncertainty in the clustering process. Clustering technology can also be used for outlier detection [21]. By the notion of rough membership function in rough set theory, Jiang et al. [22, 23] defined the rough outlier factor for outlier detection. Chen et al. [24] presented an improved clustering algorithm based on rough set and Shannon's Entropy theory.

In the procedure of clustering, neighborhood is a very important concept for describing the distribution of objects. Breuning et al. [25] introduced the concept of "Local outlier". The outlier rank of a data object is determined by taking into account the clustering structure in a bounded neighborhood of the object. Lin [26] pointed out that neighborhood spaces are more general topology space than equivalence space and introduced neighborhood relation into rough set methodology, which has shown to be a powerful tool to deal with uncertainty. Yao [27], Wu and Zhang [28] discussed the properties of neighborhood approximation spaces. Hu et al. [29] presented a conceptually simple and easy to implement method to understand and construct neighborhood-based attribute reduction technique and classifiers. In this paper, based on neighborhood-based rough set model, the cohesion degree of the neighborhood of an object and the coupling degree between neighborhoods of objects are defined, which reflect intracluster similarity and intercluster similarity, respectively. Furthermore, a new initialization method is proposed and the corresponding time complexity is analyzed as well. Finally, we compare the results of the K-Means algorithm with the proposed initialization method with that of the other two initialization methods. The experimental results show that the proposed algorithm is effective.

The outline of the rest of this paper is as follows. In Section 2, based on neighborhood-based rough set model, the cohesion degree of the neighborhood of an object and the coupling degree between neighborhoods of objects are defined. An initialization method for K-Means is proposed and the corresponding time complexity is analyzed in Section 3. In Section 4, the influence of the three norms on clustering is analyzed and the clustering results of K-Means algorithm with the three different initialization methods are compared. Then the conclusion is given in Section 5.

2. Some basic concepts

In this section, several basic concepts are reviewed, which are the neighborhood of an object, lower and upper approximations of neighborhood-based rough set model. Some novel concepts are defined, for example, the cohesion degree of the neighborhood of an object and the coupling degree between neighborhoods of objects.

2.1. Cohesion degree of the neighborhood of an object

As we know, the structural data are stored in a table, where each row (tuple) represents facts about an object. A data table is also called an information system. Data in the real world are prevalently described by numeric attributes. More formally, a numeric information system can be defined as a quadruple $IS = (U, A, V, f)$, where

U—is the nonempty set of objects, called a universe;

A—is the nonempty set of attributes;

V—the union of all attribute domains, i.e., $V = \bigcup V_a$, where V_a is the value domain of attribute a and $V \subset R$;
$f : U \times A \to V$—a mapping called an information function such that for any $x \in U$ and $a \in A$, $f(x, a) \in V_a$.

Definition 1. Let $IS = (U, A, V, f)$ be a numeric information system and $P \subseteq A$. For any $x_i, x_j \in U$, a general metric between x_i and x_j with respect to P, named Minkowski distance, is defined as

$$d_P(x_i, x_j) = \left(\sum_{m=1}^{|P|} \left(|f(x_i, a_m) - f(x_j, a_m)| \right)^{\lambda} \right)^{1/\lambda},$$

where $a_m \in P$, $\lambda = 1, 2, \infty$. When $\lambda = 1, 2, \infty$, $d_P(x_i, x_j)$ is called Manhattan distance, Euclidean distance and Chebyshev distance, respectively. For any $x_1, x_2, x_3 \in U$, it satisfies

1. $d_P(x_1, x_2) \geq 0$;
2. $d_P(x_1, x_2) = 0$ if and only if $x_1 = x_2$;
3. $d_P(x_1, x_2) = d_P(x_2, x_1)$;
4. $d_P(x_1, x_3) \leq d_P(x_1, x_2) + d_P(x_2, x_3)$.

Obviously, d_P is a distance metric.

Definition 2. Let $IS = (U, A, V, f)$ be a numeric information system and $P \subseteq A$. Given $0 < \varepsilon \leq 1$, for any $x_i \in U$, the neighborhood δ_P^{ε} of x_i with respect to P is defined as

$$\delta_P^{\varepsilon}(x_i) = \{x \in U | d_P(x, x_i) \leq \varepsilon\}.$$

When $\lambda = 1, 2, \infty$, $\delta_P^{\varepsilon}(x_i)$ is called a rhombus region, a ball region and rectangle or square around the center object x_i, respectively.

The family of neighborhood of object $\{\delta_P^{\varepsilon}(x_i) | x_i \in U\}$ covers the universe, instead of partitioning it, so we have

1. $\delta_P^{\varepsilon}(x_i) \neq \emptyset$;
2. $\bigcup_{i=1}^{|U|} \delta_P^{\varepsilon}(x_i) = U$;
3. $x_j \in \delta_P^{\varepsilon}(x_i) \Rightarrow x_i \in \delta_P^{\varepsilon}(x_j)$;
4. If $\varepsilon_1 > \varepsilon_2$, then $|\delta_P^{\varepsilon_1}(x_i)| \geq |\delta_P^{\varepsilon_2}(x_i)|$.

The size of the neighborhood depends on the threshold ε. The greater ε is, the more objects fall into the neighborhood. Here we compute ε as follows.

Since different attributes are measured on different scales, it is usual to normalize all values to lie between 0 and 1. $d_P(x_i, x_j)$ is also formulated as

$$d_P(x_i, x_j) = \left(\sum_{m=1}^{|P|} \left(\frac{|f(x_i, a_m) - f(x_j, a_m)|}{\omega} \right)^{\lambda} \right)^{1/\lambda},$$

where $\omega = \max_{i=1}^{|U|} \{f(x_i, a_m) | x_i \in U\}$. And the average distance among objects is defined as

$$\bar{x} = \frac{2}{|U|(|U| - 1)} \sum_{i=1}^{|U|-1} \sum_{j=i+1}^{|U|} d_P(x_i, x_j).$$

The size of $\bar{x}$ measures the distribution of objects in U. The greater $\bar{x}$ is, the looser distribution among objects is. Hence, in the rest of the paper, we use $\bar{x}$ to denote the size of neighborhood of objects, that is $\varepsilon = \bar{x}$.

In rough set theory [17], the upper and lower approximations are used to identify and utilize the context of each specific object and reveal relationship between objects. The upper approximation includes all objects that possibly belong to the concept while the lower approximation contains all objects that surely belong to the concept. The lower and upper approximations based on numeric data are given as follows:

Definition 3. Let $IS = (U, A, V, f)$ be a numeric information system, $P \subseteq A$ and $X \subseteq U$, the lower and upper approximations of X in U with respect to P are defined as

$$\underline{P}X = \{x_i | \delta_P^{\varepsilon}(x_i) \subseteq X, x_i \in U\},$$

and

$$\overline{P}X = \{x_i | \delta_P^{\varepsilon}(x_i) \cap X \neq \emptyset, x_i \in U\}.$$

$\underline{P}X$ is a set of objects whose neighborhood belongs to X with certainty, while $\overline{P}X$ is a set of objects whose neighborhood possibly belongs to X.

Table 1
An example data set.

Objects	a	b	Objects	a	b
x_1	1	4	x_{11}	7.4	3.9
x_2	1.3	3.9	x_{12}	7.2	3.8
x_3	1.2	3.8	x_{13}	9	4
x_4	3	4	x_{14}	8.8	3.9
x_5	2.8	3.9	x_{15}	8.9	3.7
x_6	2.9	3.7	x_{16}	8	3
x_7	2	3	x_{17}	8.1	3.1
x_8	1.9	3.3	x_{18}	4.9	1.5
x_9	2.2	3.2	x_{19}	5.4	1.7
x_{10}	8	3.2	x_{20}	5	1.8

Pawlak [17] discussed two numerical characterizations of uncertainty of a rough set: accuracy and roughness. The accuracy measures the degree of completeness of knowledge about the given rough set X, and is defined by the ratio of the cardinalities of the lower and upper approximation sets of X. Similarly, the accuracy of X in a numeric information system is formulated as

$$\alpha_P(X) = \frac{|\underline{P}X|}{|\overline{P}X|},$$

where $0 \leq \alpha_P(X) \leq 1$.

Example 1. An example data set is given by Table 1.

This is a numeric information system, where $U = \{x_1, x_2, \ldots, x_{20}\}$ and $A = \{a, b\}$. Suppose that $X = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{13}\}$ and $\lambda = 2$.

By calculating, one can have

$$\underline{A}X = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9\},$$

and

$$\overline{A}X = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{17}\}.$$

It is easy to see that

$$\alpha_A(X) = \frac{|\underline{A}X|}{|\overline{A}X|} = \frac{9}{17}.$$

$\alpha_P(X)$ measures the uncertainty with the lower bound and the upper bound of the neighborhood of an object. Suppose that there is an object which occurs in the neighborhood of each object, which means that the object which contains the maximum uncertainty provides less clustering characteristics. Based on the above ideas, the cohesion degree of the neighborhood of an object is defined as follows.

Definition 4. Let $IS = (U, A, V, f)$ be a numeric information system and $P \subseteq A$. For any $x_i \in U$, the cohesion degree of $\delta_P^\varepsilon(x_i)$ is defined as

$$Cohesion(\delta_P^\varepsilon(x_i)) = \frac{|\underline{P}(\delta_P^\varepsilon(x_i))|}{|\overline{P}(\delta_P^\varepsilon(x_i))|},$$

where $0 < Cohesion(\delta_P^\varepsilon(x_i)) \leq 1$.

The greater $Cohesion(\delta_P^\varepsilon(x))$ is, the less the boundary region of neighborhood of object x is, which means that x is a better cluster center of its neighborhood. Therefore, x is likely taken as an initial cluster center in U.

Example 2 (*Continued from* Example 1). According to Definition 4, the coupling degree of neighborhood of every object in U is shown in Table 2.

From Table 2, we have that $Cohesion(\delta_P^\varepsilon(x_6)) = Cohesion(\delta_P^\varepsilon(x_{15})) = Cohesion(\delta_P^\varepsilon(x_{18})) = Cohesion(\delta_P^\varepsilon(x_{19})) = Cohesion(\delta_P^\varepsilon(x_{20})) = 1$. However, we cannot simply take $x_6, x_{15}, x_{18}, x_{19}, x_{20}$ as initial cluster centers. This is owing to the reason that some of these objects possibly are very close and possibly belong to the same cluster. If so, the number of iteration possibly will increase and the accuracy of clustering possibly will decrease. So the coupling degree between neighborhoods of objects is given as follows.

Table 2
The coupling degree of neighborhood of every object in U ($\varepsilon = 0.3077$).

Objects	$Cohesion(\delta_P^\varepsilon(x_i))$	Objects	$Cohesion(\delta_P^\varepsilon(x_i))$
x_1	0.2222	x_{11}	0.3750
x_2	0.5556	x_{12}	0.5000
x_3	0.5556	x_{13}	0.1250
x_4	0.1111	x_{14}	0.3750
x_5	0.5556	x_{15}	1
x_6	1	x_{16}	0.1250
x_7	0.1111	x_{17}	0.2500
x_8	0.3333	x_{18}	1
x_9	0.2222	x_{19}	1
x_{10}	0.3750	x_{20}	1

2.2. Coupling degree between neighborhoods of objects

Definition 5. Let $IS = (U, A, V, f)$ be a numeric information system and $P \subseteq A$. For any $x_i, x_j \in U$, the coupling degree of $\delta_P^\varepsilon(x_i)$ and $\delta_P^\varepsilon(x_j)$ is defined as

$$Coupling(\delta_P^\varepsilon(x_i), \delta_P^\varepsilon(x_j)) = \frac{|\delta_P^\varepsilon(x_i) \cap \delta_P^\varepsilon(x_j)|}{|\delta_P^\varepsilon(x_i) \cup \delta_P^\varepsilon(x_j)|},$$

where $0 \leq Coupling(\delta_P^\varepsilon(x_i), \delta_P^\varepsilon(x_j)) \leq 1$.

The greater $Coupling(\delta_P^\varepsilon(x_i), \delta_P^\varepsilon(x_j))$ is, the more possibly x_i and x_j belong to the same cluster. In this paper, we consider that x_i and x_j belong to the same cluster, if $Coupling(\delta_P^\varepsilon(x_i), \delta_P^\varepsilon(x_j)) > \varepsilon$. On the contrary, x_i and x_j are likely taken as initial cluster centers.

The cohesion degree and the coupling degree reflect the intracluster similarity and the intercluster similarity, respectively. Based on the foregoing, an initialization method for K-Means is proposed in Section 3.

3. An initialization method for K-Means using neighborhood model

In this section, based on the cohesion degree of neighborhood of an object and the coupling degree between neighborhoods of objects, an initialization method for the K-Means algorithm is described as follows:

Input: $S = (U, A, V, f)$ and K.

Output: *Centers*.

Step 1: Initialize $Centers = \emptyset$ and $Tempcohesion = \emptyset$.

Step 2: Compute ε.

Step 3: For any $x \in U$, compute $Cohesion(\delta_A^\varepsilon(x))$. $Centers = Centers \bigcup \{x\}$ and $Tempcohesion = Tempcohesion \bigcup \{x\}$, where x satisfies $Cohesion(\delta_A^\varepsilon(x)) = \max_{i=1}^{|U|}\{Cohesion(\delta_A^\varepsilon(x_i))\}$, the first initial cluster center is selected.

Step 4: Find the next most coherent object x, where x satisfies $Cohesion(\delta_A^\varepsilon(x)) = \max\{Cohesion(\delta_A^\varepsilon(x_i))|x_i \in U\text{-}Tempcohesion\}$.

Step 5: For any $x' \in Centers$, if $Coupling(\delta_A^\varepsilon(x'), \delta_A^\varepsilon(x)) < \varepsilon$, then $Centers = Centers \bigcup \{x\}$.

Step 6: $Tempcohesion = Tempcohesion \bigcup \{x\}$.

Step 7: If $|Centers| < K$, then goto step 4, otherwise goto step 8.

Step 8: End.

The time complexity of the proposed algorithm is analyzed as follows. In Step 2, the time complexity for computing the size of neighborhood is $O(|U|^2)$. Computation of the neighborhood of objects will take $O(|U|^2)$ in Step 3. The operation on obtaining the most cohering object have a time complexity of $O(|U|)$ in Step 4. Computational cost of the rest of the steps is $O(1)$. Therefore, the entire time complexity of the proposed algorithm is $O(|U|^2)$.

4. Experimental results

In this section, experiment environments and an evaluation method [30] are introduced. Some standard data sets are downloaded. We compare the clustering results of the K-Means algorithm with the three different initialization methods, which are the proposed method, random initialization method and CCIA [15], respectively. Furthermore, the clustering results of the proposed algorithm with the three different norms are analyzed as well. The experimental results show that the proposed method outperforms the other two initialization methods.

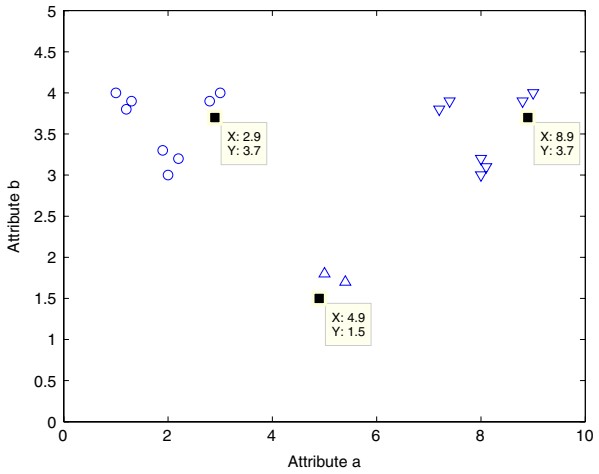

Fig. 1. Clustering result of the K-Means algorithm with the proposed initialization method on Example data ($K = 3$).

4.1. Experimental environments and evaluation method

The experiments are conducted on a PC with an Intel Pentium 4 processor (2.4 GHz) and 1G byte memory running the Windows XP SP3 operating system. The K-Means algorithms with the three different initialization methods are coded in MATLAB 7.0 programming language.

To evaluate the efficiency of clustering algorithms, three evaluation index accuracy (AC), precision (PR), and recall (RE) are employed in the following experiments. In order to define the three kinds of evaluation indexes, the following quantities are needed.

k—the number of classes of the data, which is known;
a_i—the number of objects that are correctly assigned to the class C_i $(1 \leq i \leq k)$;
b_i—the number of objects that are incorrectly assigned to the class C_i;
c_i—the number of objects that should be in, but are not correctly assigned to the class C_i;

The accuracy, precision and recall are defined as $AC = \frac{\sum_{i=1}^{k} a_i}{|U|}$, $PR = \frac{\sum_{i=1}^{k}(\frac{a_i}{a_i+b_i})}{k}$, $RE = \frac{\sum_{i=1}^{k}(\frac{a_i}{a_i+c_i})}{k}$, respectively.

4.2. Evaluation on clustering effectiveness

In order to test the effectiveness of the proposed method, some data sets are downloaded from the machine learning data repository, University of California at Irvine [31]. Some experiments are done on these data sets and the results are compared with the other two initialization methods, which are random initialization and CCIA [15]. As the K-Means algorithm is especially sensitive to initial cluster centers, we carry out 100 runs of the k-Means algorithm with random initialization methods on these standard data sets, respectively. Therefore, we take the average of 100 times experiments as experimental results. The comparison results of the k-Means algorithm with three initialization methods on Example data set and real data sets are shown, respectively.

Example data (Table 1)

Suppose that the desired number of clusters is 3 or 7, thus the corresponding initial cluster centers $\{x_6, x_{15}, x_{18}\}$ and $\{x_6, x_{15}, x_{18}, x_2, x_{12}, x_8, x_{17}\}$ can be obtained by the proposed algorithm. We take $\{x_6, x_{15}, x_{18}\}$ and $\{x_6, x_{15}, x_{18}, x_2, x_{12}, x_8, x_{17}\}$ as the initial cluster centers of the K-Means algorithm, respectively. Figs. 1 and 2 show the results of the K-Means algorithm with $K = 3$ or $K = 7$, respectively. Note that objects with black square in Figs. 1 and 2 are taken as initial cluster centers in clustering process.

From Figs. 1 and 2, we find that the proposed algorithm can accurately discover the initial cluster centers for the K-Means algorithm.

Furthermore, Tables 3 and 4 illustrate the influence of the three norms on clustering and the comparison results of three initialization methods on Example data with $K = 3$ or $K = 7$, respectively.

Table 3 shows that the proposed method and CCIA outperform random initialization method. However, in Table 4, the results of CCIA is much lower than that of the proposed method and random initialization. The reasons are listed below. In Table 1, for attribute a and b, CCIA generates five class label, respectively. Thus, we obtain the number of the unique strings, which is less than $K = 7$. Therefore, CCIA cannot be conducted correctly.

Iris data

This data set has often been used as a standard for testing clustering algorithms. This data set has three classes that represent three different varieties of Iris flowers namely Iris setosa, Iris versicolor and Iris virginica. Fifty objects are obtained

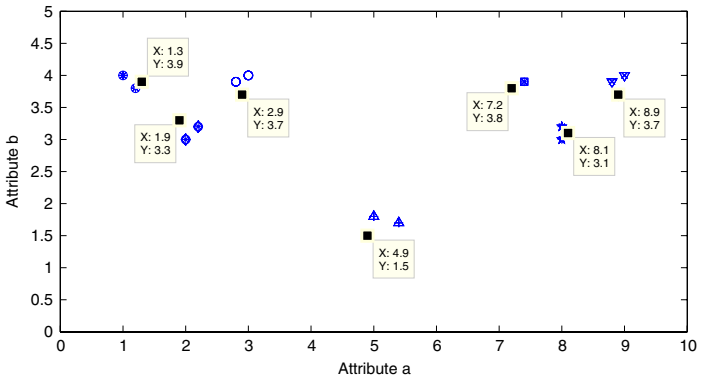

Fig. 2. Clustering result of the K-Means algorithm with the proposed initialization method on Example data ($K = 7$).

Table 3
The comparison results of the K-Means algorithm with the three initialization methods on Example data ($K = 3$).

	The proposed method			Random method	CCIA
	$\lambda = 1$	$\lambda = 2$	$\lambda = \infty$		
AC	1	1	1	0.9625	1
PR	1	1	1	0.9671	1
RE	1	1	1	0.9713	1

Table 4
The comparison results of the K-Means algorithm with the three initialization methods on Example data ($K = 7$).

	The proposed method			Random method	CCIA
	$\lambda = 1$	$\lambda = 2$	$\lambda = \infty$		
AC	0.8500	1	0.8500	0.8120	0.4500
PR	0.7857	1	0.7857	0.7074	0.2440
RE	1	1	1	0.9886	1

Table 5
The comparison results of the K-Means algorithm with the three initialization methods on Iris data.

	The proposed method			Random method	CCIA
	$\lambda = 1$	$\lambda = 2$	$\lambda = \infty$		
AC	0.8867	0.6667	0.6667	0.8452	0.8867
PR	0.8868	0.7797	0.7797	0.8809	0.8979
RE	0.8867	0.8533	0.8533	0.8786	0.8867

Table 6
The comparison results of the K-Means algorithm with the three initialization methods on Wine data.

	The proposed method			Random method	CCIA
	$\lambda = 1$	$\lambda = 2$	$\lambda = \infty$		
AC	0.9494	0.9438	0.9494	0.9442	0.9438
PR	0.9496	0.9412	0.9496	0.9440	0.9412
RE	0.9577	0.9521	0.9577	0.9547	0.9521

from each of the three classes, thus a total of 150 objects is available. Every object is described by four attributes, viz sepal length, sepal width, petal length and petal width. The experimental results are summarized in Table 5.

Wine recognition data

This data set is the result of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines. There are overall 178 objects. There are 59, 71, 48 objects in class I, class II and class III respectively. The experimental results are summarized in Table 6.

Glass data

This data set has 214 objects and 10 (including an Id#) attributes. There are 7 clusters (70 building windows, 17 vehicle windows, 76 building windows, 0 vehicle windows, 13 containers, 9 tableware and 29 headlamps) that can be grouped in 2

Table 7
The comparison results of the K-Means algorithm with the three initialization methods on Glass data.

	The proposed method			Random method	CCIA
	$\lambda = 1$	$\lambda = 2$	$\lambda = \infty$		
AC	0.8972	0.8972	0.8972	0.8779	0.7617
PR	0.8584	0.8584	0.8584	0.8361	0.8257
RE	0.8584	0.8584	0.8584	0.8467	0.8086

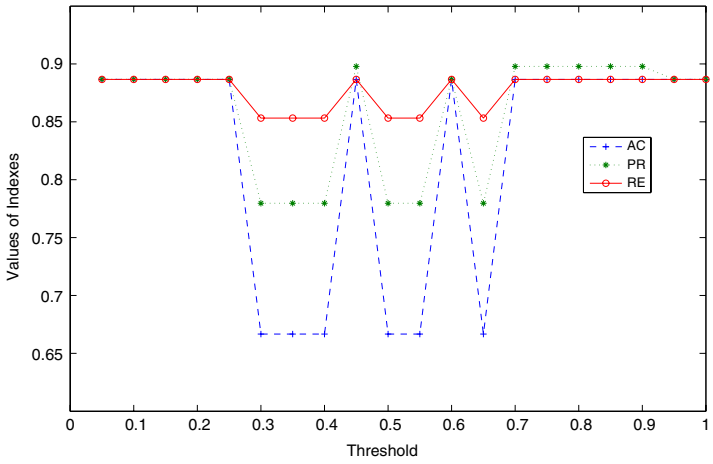

Fig. 3. Evaluation index curves varying with ε on Iris data.

bigger clusters (163 Window glass, 51 Non-window glass). In this experiment, suppose that the number of clusters is 2. The experimental results are summarized in Table 7.

From Tables 5–7, we can find that the clustering results of the K-Means algorithms with the three initialization methods are very close on Wine data set. The proposed method with $\lambda = 1$ and CCIA are superior to random initialization method on Iris data set. On Glass data set, the proposed method outperforms the other two initialization methods. Furthermore, random initialization method is superior to CCIA. In addition, the clustering results on the most data sets are insensitive to norms, except for Iris data set.

4.3. Sensitivity analysis

As the size of neighborhood has direct influence to clustering results, we conduct a series of experiments to find the optimal parameter ε used to control the size of the neighborhood. We try ε from 0.05 to 1 with step 0.05, and present the clustering results of the K-Means algorithm with the proposed method on real data sets, respectively, where $\lambda = 2$. Figs. 3–5 present the three evaluation index curves varying with ε for different data sets: Iris, Wine and Glass.

From Fig. 3, we can find that the threshold ε is in range $[0.05, 0.25]$ or $[0.7, 1]$, which is optimal or near optimal on Iris data set. In Fig. 5, it is clear that the threshold ε is in range $[0.05, 0.2]$, which is near optimal. On Wine data set, the clustering results are optimal, when ε is between 0.35 and 0.4. Therefore, we recommend that ε should take values in the range $[0.1, 0.2]$.

5. Conclusions

K-Means algorithm is widely discussed and applied, however, the K-Means algorithm suffers from initial starting condition effects. As a computing model of information granular, neighborhood model has been successfully applied. In this paper, the cohesion degree of the neighborhood of an object and the coupling degree between neighborhoods of objects have been defined based on the neighborhood-based rough set model. Furthermore, a new initialization method has been proposed, and the corresponding time complexity has been analyzed as well. We study the influence of the three norms on clustering, and compare the proposed initialization method with the other two initialization methods. The experimental results illustrate the effectiveness of the proposed method.

Acknowledgements

The authors are very grateful to the anonymous reviewers and editor. Their many helpful and constructive comments and suggestions helped us significantly improve this work. This work was also supported by the National Natural Science

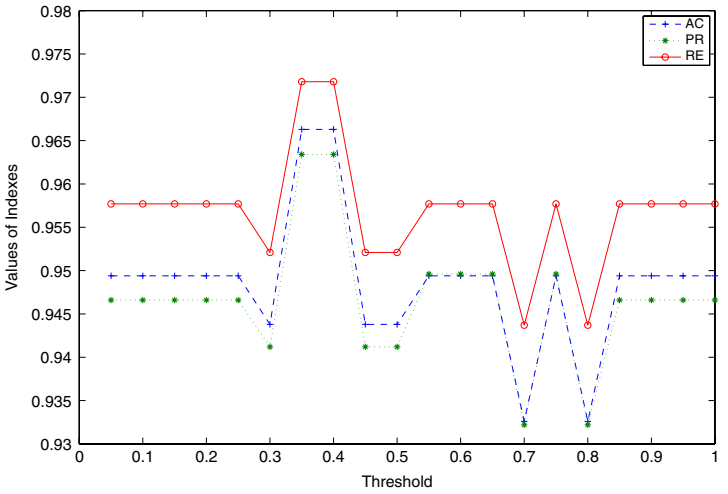

Fig. 4. Evaluation index curves varying with ε on Wine data.

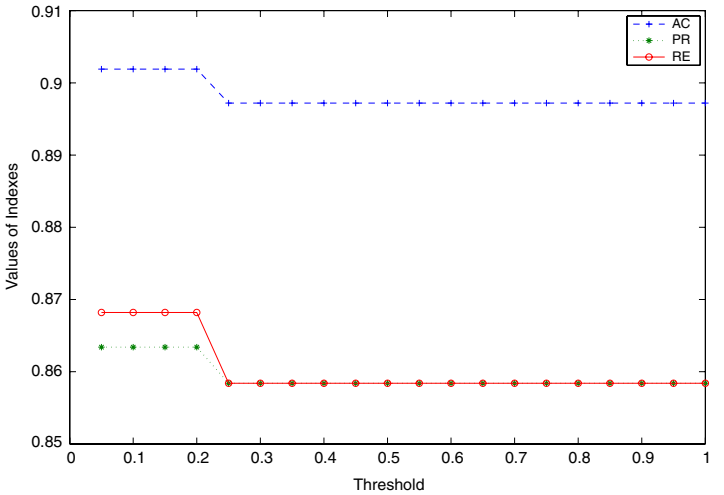

Fig. 5. Evaluation index curves varying with ε on Glass data.

Foundation of China (Nos. 60773133, 60573074, 60875040), the High Technology Research and Development Program of China (863) (No. 2007AA01Z165), the National Key Basic Research and Development Program of China (973) (No. 2007CB311002), the Doctor Authorization Foundation of the Ministry of Education (No. 200801080006), the Natural Science Foundation of Shanxi (No. 2008011038), the Key Laboratory Open Foundation of Shanxi (Nos. 200603023, 2007031017) and the Technology Research Development Projects of Shanxi (No. 2007103).

References

[1] J. Han, M. Kamber, Data Mining Concepts and Techniques, Morgan Kaufmann, San Francisco, 2001.
[2] J.F. Brendan, D. Delbert, Clustering by passing messages between data points, Science 315 (16) (2007) 972–976.
[3] Q.J. Mac, Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium, vol. 1, 1967, pp. 281–297.
[4] J.M. Pen, J.A. Lozano, P. Larraaga, An empirical comparison of four initialization methods for the *K*-Means algorithm, Pattern Recognition Letter 20 (1999) 1027–1040.
[5] S.Z. Selim, M.A. Ismail, *K*-Means-type algorithms: a generalized convergence theorem and characterization of local optimality, IIEEE Transactions on Pattern Analysis and Machine Intelligence 6 (1984) 81–87.
[6] R.O. Duda, P.E. Hart, Pattern Classification and Scene Analysis, John Wiley and Sons, NY, 1973.
[7] G.W. Milligan, An examination of the eect of six types of error perturbation on Rfteen clustering algorithms, Psychometrika 45 (1980) 325–342.
[8] J.H. Ward, Hierarchical grouping to optimize an objective function, Journal of the American Statistical Association 58 (1963) 236–244.
[9] D. Fisher, Iterative optimization and simplification of hierarchical clusterings, Journal of Artificial Intelligence Research 4 (1996) 147–179.
[10] D. Fisher, Knowledge acquisition via incremental conceptual clustering, Machine Learning 2 (1987) 139–172.
[11] R.E. Higgs, K.G. Bemis, I.A. Watson, J.H. Wikel, Experimental designs for selecting molecules from large chemical databases, Journal of Chemical Information and Computer Sciences 37 (1997) 861–870.

[12] M. Snarey, N.K. Terrett, P. Willet, D.J. Wilton, Comparison of algorithms for dissimilarity-based compound selection, Journal of Molecular Graphics and Modelling 15 (1997) 372–385.
[13] P.S. Bradley, O.L. Mangasarian, W.N. Street, Clustering via concave minimization, in: M.C. Mozer, M.I. Jordan, T. Petsche (Eds.), in: Advances in Neural Information Processing System, vol. 9, MIT Press, 1997, pp. 368–374.
[14] P.S. Bradley, U.M. Fayyad, Refining initial points for *K*-Means clustering, in: J. Sharlik (Ed.), Proc. 15th International Conference on Machine Learning (ICML 98), Morgan Kaufmann, San Francisco, CA, 1998, pp. 91–99.
[15] S.S. Khan, A. Ahmad, Cluster center initialization algorithm for *K*-Means clustering, Patter Recognition Letters 25 (2004) 1293–1302.
[16] M. Meila, D. Heckerman, An experimental comparison of several clustering and initialization methods, in: Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann, San Francisco, CA, 1998, pp. 386–395.
[17] Z. Pawlak, Rough Sets-Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, Dordrecht, Boston, London, 1991.
[18] J.Y. Liang, D.Y. Li, Uncertainty and Knowledge Acquisition in Information Systems, Science Press, Beijing, 2005.
[19] J.Y. Liang, J.H. Wang, Y.H. Qian, A new measure of uncertainty based on knowledge granulation for rough sets, Information Sciences 179 (4) (2009) 458–470.
[20] D. Parmar, T. Wu, J. Blackhurst, MMR: An algorithm for clustering data using rough set theory, Data & Knowledge Engineering 63 (3) (2007) 893–897.
[21] A.K. Jain, M.N. Murty, P.J. Flynn, Data clustering: A review, ACM Computing Surveys 31 (3) (1999) 264–323.
[22] F. Jiang, Y.F. Sui, C.G. Cao, A rough set approach to outlier detection, International Journal of General Systems 37 (5) (2008) 519–536.
[23] F. Jiang, Y.F. Sui, C.G. Cao, Some issues about outlier detection in rough set theory, Expert Systems with Applications 36 (2009) 4680–4687.
[24] C.B. Chen, L.Y. Wang, Rough set-based clustering with refinement using Shannon's Entropy theory, Computer and Mathematics with Applications 52 (2006) 1563–1576.
[25] M.M. Breunig, H.P. Kriegel, R.T. Ng, J. Sander, LOF: Identifying density-based local outliers, in: Proceedings of SIGMOD, Dallas, Texas, 2000, 427-438.
[26] T.Y. Lin, Granular Computing on binary relations I: Data mining and neighborhood systems, in: A Skoworn, L Pokowshi (Eds.), rough sets in knowledge discovery, Physica-Verlag, 1998, pp. 107–121.
[27] Y.Y. Yao, Relational interpretation of neighborhood operators and neighborhood systems, Information Sciences 111 (198) (1998) 239–259.
[28] W.Z. Wu, W.X. Zhang, Neighborhood operator systems and approximations, Information Sciences 144 (1–4) (2002) 201–217.
[29] Q.H. Hu, D.R. Yu, Z.X. Xie, Neighborhood classifiers, Expert Systems with Applications 34 (2) (2008) 866–876.
[30] Y.M. Yang, An evaluation of statistical approaches to text categorization, Journal of Information Retrieval 1 (1-2) (1999) 67–88.
[31] The UCI Machine Learning Repository http://mlearn.ics.uci.edu/MLRepository.html.