Expert Systems with Applications 39 (2012) 8022–8029

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

A cluster centers initialization method for clustering categorical data

Liang Bai [a,b], Jiye Liang [a,*], Chuangyin Dang [b], Fuyuan Cao [a]

[a] *Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, School of Computer and Information Technology, Shanxi University, Taiyuan 030006, Shanxi, China*

[b] *Department of Manufacturing Engineering and Engineering Management, City University of Hong Kong, Hong Kong*

ARTICLE INFO

Keywords:
The *k*-modes algorithm
Initialization method
Initial cluster centers
Density
Distance

ABSTRACT

The leading partitional clustering technique, *k*-modes, is one of the most computationally efficient clustering methods for categorical data. However, the performance of the *k*-modes clustering algorithm which converges to numerous local minima strongly depends on initial cluster centers. Currently, most methods of initialization cluster centers are mainly for numerical data. Due to lack of geometry for the categorical data, these methods used in cluster centers initialization for numerical data are not applicable to categorical data. This paper proposes a novel initialization method for categorical data which is implemented to the *k*-modes algorithm. The method integrates the distance and the density together to select initial cluster centers and overcomes shortcomings of the existing initialization methods for categorical data. Experimental results illustrate the proposed initialization method is effective and can be applied to large data sets for its linear time complexity with respect to the number of data objects.

1. Introduction

Clustering is a process of grouping a set of objects into clusters so that the objects in the same cluster have high similarity but are very dissimilar with objects in other clusters. Various types of clustering methods have been proposed and developed, see, for instance (Jain & Dubes, 1988). Clustering algorithms in the literature can generally be classified into two types: hierarchical clustering and partitional clustering. Hierarchical clustering algorithms, essentially heuristic procedures, produce a hierarchy of partitions of the set of observations according to an agglomerative strategy or to a divisive one. Partitional clustering algorithms, in general, assume a given number of clusters and, essentially, seek the optimization of an objective function measuring the homogeneity within the clusters and/or the separation between the clusters.

The *k*-means algorithm (Anderberg, 1973; Ball & Hall, 1967; MacQueen, 1967; Jain & Dubes, 1988) is a well known partitional clustering algorithm which is widely used in real world applications such as marketing research and data mining to cluster very large data sets due to their efficiency. In 1997 Huang (1997, 1998), extended the *k*-means algorithm to propose the *k*-modes algorithm whose extensions have removed the numeric-only limitation of the *k*-means algorithm and enable the *k*-means clustering process to be used to efficiently cluster large categorical data sets from real world databases. Since first published, the *k*-modes algorithm has become a popular technique in solving categorical data clustering problems in different application domains (Andreopoulos, An, & Wang, 2005).

The *k*-means algorithm and the *k*-modes algorithm use alternating minimization methods to solve non convex optimization problems in finding cluster solutions (Jain & Dubes, 1988). These algorithms require a set of initial cluster centers to start and often end up with different clustering results from different sets of initial cluster centers. Therefore, these algorithms are very sensitive to the initial cluster centers. Usually, these algorithms are run with different initial guesses of cluster centers, and the results are compared in order to determine the best clustering results. One way is to select the clustering results with the least objective function value formulated in these algorithms, see, for instance (Huang, Ng, Rong, & Li, 2005). In addition, cluster validation techniques can be employed to select the best clustering result, see, for instance (Jain & Dubes, 1988). Other approaches have been proposed and studied to address this issue by using a better initial seed value selection for the *k*-means algorithm (Arthur & Vassilvitskii, 2007; Babu & Murty, 1993; Brendan & Delbert, 2007; Bradley, Mangasarian, & Street, 1997; Bradley & Fayyad, 1998; Khan & Ahmad, 2004; Krishna & Murty, 1999; Laszlo & Mukherjee, 2006, 2007; Pen, Lozano, & Larraaga, 1999). For example, some experts (Babu & Murty, 1993; Krishna & Murty, 1999; Laszlo & Mukherjee, 2006, 2007) used genetic algorithm to obtain the good initial cluster centers. Arthur and Vassilvitskii (2007) proposed and studied a careful seeding for initial cluster centers to improve clustering results. However, due to lack of intuitive geometry for categorical data, the techniques used in cluster centers initialization for numerical data are not applicable to categorical data. To date, few researches

[*] Corresponding author.

E-mail addresses: sxbailiang@126.com (L. Bai), ljy@sxu.edu.cn (J. Liang), mecdang@cityu.edu.hk (C. Dang), cfy@sxu.edu.cn (F. Cao).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2012.01.131

are concerned for cluster centers initialization for categorical data. However, due to the fact that large categorical data sets exist in many applications, it has been widely recognized that directly clustering the raw categorical data is important. Examples include environmental data analysis (Wrigley, 1985), market basket data analysis (Aggarwal, Magdalena, & Yu, 2002), DNA or protein sequence analysis (Baxevanis & Ouellette, 2001), text mining (Wang & Karypis, 2006), and computer security (Barbara & Jajodia, 2002). Therefore, how to select initial cluster centers for clustering categorical data become an important research question. The *k*-centers clustering technique.

Huang in Huang (1998) suggested to select the first *k* distinct objects from the data set as the initial *k* modes or assign the most frequent categories equally to the initial *k* modes. Though the methods are to make the initial modes diverse, an uniform criteria is not given for selecting *k* initial modes in Huang (1998). Sun, Zhu, and Chen (2002) introduces an initialization method which is based on the frame of refining. This method presents a study on applying Bradley's iterative initial-point refinement algorithm (Bradley & Fayyad, 1998) to the *k*-modes clustering, but its time cost is high and the parameters of this method are plenty which need to be asserted in advance. In Coolcat algorithm (Barbara, Couto, & Li, 2002), the MaxMin distances method is used to find the *k* most dissimilar data objects from the data set as initial seeds. However, the method only considers the distance between the data objects, by which outliers maybe be selected. Cao, Liang, and Bai (2009) and Wu, Jiang, and Huang (2007) integrated the distance and the density together to propose a cluster centers initialization method, respectively. The difference between the two methods is the definition of the density of an object. Wu used the total distance between an object and all objects from data set as the density of the object. Due to the fact that the time complexity of calculating the densities of all objects is $O(n^2)$, it limits the process in a sub-sample data set and uses a refining framework. But this method needs to randomly select sub-sample, so the sole clustering result can not be guaranteed. Cao et al. (2009) defined the density of an object based on frequency of attribute values. In this paper, we prove that Cao's density is equivalent to Wu's density, which means that Cao's method is equivalent to Wu's method. Although the two methods can avoid to select outliers as the cluster centers by the density, they have some shortcomings: (1) The object with the maximum density is taken as the first cluster center. Due to the fact that they only considered the factor of density in the selection of the first cluster center, it is possible that the selected object is a boundary point among clusters, which is proved in this paper; (2) One real object in a cluster is selected as the cluster center. But in most cases, the center of a cluster is not a real object but a virtual object, which means that a real object could not sufficiently represent the cluster. In summary, there are no universally accepted method for obtaining initial cluster centers currently. Hence, it is very necessary to propose a new initialization method for categorical data which overcomes shortcomings of the existing initialization methods.

In the paper, we propose a novel cluster centers initialization method. We use the distances between objects and the center of the whole data set to avoid selecting the boundary objects among clusters as the first cluster center. In this method, an object is selected not as an initial cluster center but as a cluster exemplar. We integrate the cluster exemplar and the neighbor objects around it together to construct the candidates of the initial cluster center, and define some criteria to select initial cluster centers from the candidates. The proposed initialization method is used along with the *k*-modes algorithm. The time complexity of the method is analyzed. The comparisons with other methods illustrate the effectiveness of this approach.

The outline of the rest of this paper is as follows. Section 2 introduces the *k*-modes algorithm. In Section 3, a new initialization method is proposed. Section 4 demonstrates the effectiveness and scalability of the new initialization method. General discussion and the conclusions of this work follow in Section 5.

2. The *k*-modes algorithm

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 in rough set theory (Liang & Li, 2005, Liang, Wang, & Qian, 2009; Pawlak, 1991; Qian, Liang, Pedrycz, & Dang, 2010). Data in the real world usually contains categorical attributes (Gowda & Diday, 1999). More formally, a categorical data table is defined as a quadruple $IS = (U,A,V,f)$, where:

(1) U is the nonempty set of objects, called a universe.
(2) A is the nonempty set of attributes.
(3) V is the union of attribute domains, i.e., $V = \bigcup_{a\in A} V_a$, where V_a is the value domain of attribute a and it is finite and unordered, e.g., for any p, $q \in V_a$, either $p = q$ or $p \neq q$.
(4) $f{:}U \times A \to V$ is an information function such that for any $a \in A$ and $x \in U$, $f(x,a) \in V_a$.

The objective of the *k*-modes algorithm is to cluster U into k clusters by minimizing the function

$$F(W,Z) = \sum_{l=1}^{k}\sum_{i=1}^{n}\omega_{li} d(z_l, x_i)$$

subject to

$$\omega_{li} \in \{0,1\}, \quad 1 \leqslant l \leqslant k, \ \ 1 \leqslant i \leqslant n,$$

$$\sum_{l=1}^{k}\omega_{li} = 1, \quad 1 \leqslant i \leqslant n,$$

and

$$0 < \sum_{i=1}^{n}\omega_{li} < n, \quad 1 \leqslant l \leqslant k,$$

where $k(\leqslant n)$ is a known number of clusters. $W = [\omega_{li}]$ is a k-by-n $\{0,1\}$ matrix, ω_{li} indicates whether x_i belongs to the lth cluster for the *k*-modes algorithm, $\omega_{li} = 1$ if x_i belongs to the lth cluster and 0 otherwise. $Z = \{z_1, z_2, \ldots, z_k\}$, and z_l is the lth cluster center with the categorical attributes $a_1, a_2, \ldots, a_{|A|}$.

To cluster categorical data, the *k*-modes algorithm (Huang, 1997, 1998) measures the distance between a cluster center and a categorical data object, and updates the cluster center at each iteration as follows:

The distance measure $d(z_l,x_i)$ between a center z_l and a categorical data object x_i is defined as

$$d(z_l, x_i) = \sum_{a\in A}\delta_a(z_l, x_i),$$

where

$$\delta_a(z_l, x_i) = \begin{cases} 1, & f(z_l, a) \neq f(x_i, a), \\ 0, & f(z_l, a) = f(x_i, a). \end{cases}$$

It is easy to verify that the function d defines a metric space on the set of categorical objects. The lth cluster center z_l, referred to as the lth mode, is updated as follows. Each $f(z_l,a)$ for $a \in A$ is updated. For the *k*-modes algorithm, $f(z_l,a)$ satisfies the following criterion:

$$|\{x_i \in U | f(x_i, a) = f(z_l, a), \omega_{li} = 1\}| = \max_{q\in V_a} |\{x_i \in U | f(x_i, a) = q, \omega_{li} = 1\}|.$$

For the *k*-modes algorithm, $W = [\omega_{li}]$ is updated as

$$\omega_{li} = \begin{cases} 1, & \text{if} \quad d(z_l, x_i) = \min\limits_{1\leqslant h\leqslant k} d(z_h, x_i), \\ 0, & \text{otherwise}. \end{cases}$$

The whole process of the k-modes algorithm is described as follows (Huang, 1998):

Step 1. Choose an initial point $Z^{(1)} \subseteq R$, where $R = V_{a_1} \times V_{a_2} \times \cdots \times V_{a_{|A|}}$. Determine $W^{(1)}$ such that $F(W,Z^{(1)})$ is minimized. Set $t = 1$.
Step 2. Determine $Z^{(t+1)}$ such that $F(W^{(t)}, Z^{(t+1)})$ is minimized. If $F(W^{(t)}, Z^{(t+1)}) = F(W^{(t)}, Z^{(t)})$, then stop; otherwise goto Step 3.
Step 3. Determine $W^{(t+1)}$ such that $F(W^{(t+1)}, Z^{(t+1)})$ is minimized. If $F(W^{(t+1)}, Z^{(t+1)}) = F(W^{(t)}, Z^{(t+1)})$, then stop; otherwise set $t = t + 1$ and goto Step 2.

This procedure removes the numeric-only limitation of the k-means algorithm. However, the k-modes algorithm is the same as the k-means algorithm and is sensitive to initial cluster centers. To solve these problem, a new initialization method for categorical data is proposed in Section 3.

3. A new cluster centers initialization method

Currently, many approaches are proposed to measure the cohesiveness of a data object (Cao et al., 2009; Ester, Kriegel, Sander, & Xu, 1996; Wu et al., 2007). It is well known to use the total distance between a data object and all data objects to measure the density of the data object (Wu et al., 2007), because of its simple and no parameters. Hence, we decide to use the density to measure a point in R where $R = V_{a_1} \times V_{a_2} \times \cdots \times V_{a_{|A|}}$. The definition is given as follows:

Definition 1. Let $IS = (U,A,V,f)$ be a categorical data table, $A = \{a_1, a_2, \ldots, a_{|A|}\}$, $R = V_{a_1} \times V_{a_2} \times \cdots \times V_{a_{|A|}}$. For any $x \in R$, the density of x in R is defined as

$$Dens(x) = -\frac{1}{|U|}\sum_{y \in U} d(x,y).$$

Obviously, we have $-|A| \leqslant Dens(x) \leqslant 0$. For any $y \in U$, if $d(x,y) = 0$, then $Dens(x) = 0$. If $d(x,y) = |A|$, then $Dens(x) = -|A|$.

By the definition of the density, we know that the time complexity of calculating the densities of n objects is $O(|U||A|n)$. Since the initialization method needs to calculate the densities of all data objects in the data set whose time complexity is $O(|U|^2|A|)$, a working method has to optimize the density calculation.

Proposition 1. *For any point* $x \in R$, $Dens(x) = \sum_{a \in A}\left(\frac{|\{y \in U | f(x,a) = f(y,a)\}|}{|U|} - 1\right)$.

Proof

$$\begin{aligned} Dens(x) &= -\frac{1}{|U|}\sum_{y \in U} d(x,y) = -\frac{1}{|U|}\sum_{y \in U}\sum_{a \in A}\delta_a(x,y) \\ &= -\frac{1}{|U|}\sum_{a \in A}\sum_{y \in U}\delta_a(x,y) \\ &= -\frac{1}{|U|}\sum_{a \in A}(|U| - |\{y \in U | f(x,a) = f(y,a)\}|) \\ &= \sum_{a \in A}\left(\frac{|\{y \in U | f(x,a) = f(y,a)\}|}{|U|} - 1\right). \end{aligned}$$

This completes the proof. □

From the view of Proposition 1, we know the relation between the density and the frequencies of the attribute values, which prove that Cao's density is equivalent to Wu's density. For a given categorical data table, the number of every category of every categorical attribute is known. We first compute the frequency of every attribute value of every attribute and save these to a table. Next, we compute densities of data objects by the saved table. The method is described in Table 1. Since the time complexity of calculating frequencies of all values in the categorical attribute $a \in A$ is $O(|U||V_a|)$, the time complexity of calculating densities of all data objects in a data table is $O(|U||V|)$, where $|V| = \sum_{a \in A}|V_a|$. When the number of objects is large, $|V| \ll |U|$. Therefore, the method in Table 1 can be applied to large data sets for its linear time complexity with respect to the number of data objects.

Proposition 2. *If $z \in R$ is a mode of U, then $Dens(z) = \max_{x \in R} Dens(x)$.*

Proof. Let z be a mode of U. From the definition of mode in Section 2, it follows that for each $a \in A$,

$$|\{x \in U | f(x,a) = f(z,a)\}| = \max_{q \in V_a}|\{x \in U | f(x,a) = q\}|,$$

then

$$Dens(z) = \max_{x \in R} Dens(x).$$

This completes the proof. □

According to Definition 1, we know that the more $Dens(x)$ is, if can be expressed in a graph, the more the number of objects around x is, as well as the more possible x be a cluster center. So Wu et al. (2007) and Cao et al. (2009) selected the point with the maximum density as the first initial cluster center. However, the point also may be a boundary point among clusters. According to Proposition 2, we find that the $Dens(z)$ of the mode z of U is maximum. z can be seen as a center point of U which is similar to the mean of numerical data and reflects the common features in U. That means that the smaller the distance between a data object x and z is, the more x maybe contain the common features of all clusters. When a data object x contains many common features of all clusters, x may be a boundary point among clusters although x maybe have high value of $Dens(x)$.

Let us consider the following example to demonstrate the problem. The synthetic data set in Table 2 is described with four categorical attributes a_1 (four categories: B, C, E or F), a_2 (six categories: B, C, D, E, F or G), a_3 (five categories: B, C, D, E or F) and a_4(six categories: B, C, D, E or F), and there are three classes with with their modes and their four objects.

We find that x_4 is a boundary point among clusters, although $Dens(x_4) = \max_{i=1}^{12} Dens(x_i)$. Since $d(x_4, z_1) = \max_{y \in D_1} d(y, z_1)$, x_4 cannot better reflect the characteristics of the class D_1 than other objects in D_1.

Table 1
Computation of the densities of all objects in U.

```
Input: $IS = (U,A,V,f)$ and $k$, where $k$ is the number of cluster desired.
Output: $Dens(x)$, $x \in U$.
Let $fr_{a,q}$ denotes the number of categorical objects in $U$ which have the value $q$
   of the attribute $a$, $q \in V_a$, $a \in A$.
Begin
  $Centers = \emptyset$;
  $fr_{a,q} = 0$, $1 \leqslant i \leqslant |A|$, $q \in V_a$, $a \in A$;
  For each $x$ in $U$
    For each $a$ in $A$
      For each $q$ in $V_a$
        if $f(y,a) == q$
          $fr_{a,q} = fr_{a,q} + 1$;
        end
      end
    end
  end
  For each $x$ in $U$
    Compute $Dens(x)$ using Proposition 1;
  end
end
```

Table 2
Synthetic data set.

Objects	Attributes				
	a_1	a_2	a_3	a_4	Class
x_1	*B*	*B*	*F*	*B*	D_1
x_2	*B*	*F*	*B*	*B*	D_1
x_3	*B*	*B*	*B*	*E*	D_1
x_4	*C*	*E*	*B*	*B*	D_1
The mode z_1 of the class D_1	*B*	*B*	*B*	*B*	
x_5	*C*	*C*	*D*	*C*	D_2
x_6	*C*	*C*	*C*	*D*	D_2
x_7	*C*	*D*	*C*	*C*	D_2
x_8	*E*	*G*	*C*	*C*	D_2
The mode z_2 of the class D_2	*C*	*C*	*C*	*C*	
x_9	*E*	*E*	*B*	*E*	D_3
x_{10}	*F*	*E*	*E*	*E*	D_3
x_{11}	*E*	*E*	*E*	*F*	D_3
x_{12}	*C*	*B*	*E*	*E*	D_3
The mode z_3 of the class D_3	*E*	*E*	*E*	*E*	

Table 3
The method for computing CA_x.

```
Input: $IS = (U,A,V,f)$ and $x \in U$.
Output: $CA_x$.
Let $S_i = \{y|d(x,y) = i, y \in U\}$ and $Q_i = \{y|d(x,y) \leqslant i, y \in U\}$, $1 \leqslant i \leqslant |A|$. $fr_{i,a,q}$
    denotes the number of categorical objects in $S_i$ which have the value $q$ of
    the attribute $a$ and $qfr_{i,a,q}$ denotes the number of categorical objects in $Q_i$
    which have the value $q$ of the attribute $a$, $1 \leqslant i \leqslant |A|$, $q \in V_a$, $a \in A$.
Begin
  $CA_x = \emptyset$;
  $S_i = \emptyset$, $1 \leqslant i \leqslant |A|$;
  $fr_{i,a,q} = 0$, $1 \leqslant i \leqslant |A|$, $q \in V_a, a \in A$;
  For each $y$ in $U$
    $i = d(x,y)$;
    $S_i = S_i \bigcup \{y\}$;
    For each $a$ in $A$
      For each $q$ in $V_a$
        if $f(y,a) = = q$
          $fr_{i,a,q} = fr_{i,a,q} + 1$;
        end
      end
    end
  end
  For $i = 1$ to $|A|$
    if $i = =1$
      $Q_i = S_i$;
      $qfr_{i,a,q} = fr_{i,a,q}, q \in V_a, a \in A$;
    else
      $Q_i = Q_{i-1} + S_i$;
      $qfr_{i,a,q} = qfr_{i-1,a,q} + fr_{i,a,q}, q \in V_a, a \in A$;
    end
    Compute $ca_i$ where is a mode of $Q_i$;
    $CA_x = ca_i \bigcup CA_x$;
  end
end
```

Therefore, when computing the first cluster center, if the density of the objects is only taken into account, it is very possible that the boundary point among clusters is taken as a cluster center. To avoid the potential problem, first, we compute a mode z of U. Since z reflects the common features in U, if the smaller the distances between an object and z is, the more possible the object as a boundary point among clusters is. So we select the object with higher density and farther from z. We combine the distance between the object and z with the density of the object together to measure the possibility of the object to be an exemplar of the first cluster. According to Definition 2, we select a data object with maximum $Pos\_Exemplar_{C_1}(x)$ as an exemplar of the first cluster.

Definition 2. Let $IS = (U,A,V,f)$ be a categorical data table and $z \in R$ be a mode of U. For any $x \in U$, the possibility of x to be an exemplar of the first cluster C_1 is defined as

$$Pos\_Exemplar_{C_1}(x) = Dens(x) + d(x,z).$$

Due to the fact that an object could not sufficiently represent a cluster, we take the selected object not as an initial cluster center but as an exemplar of a cluster. We integrate the information of the exemplar and its neighbor objects to construct the candidates and define some criteria to select points from the candidates as the initial cluster centers. In the following, we give the definition and method of constructing the candidates based on the selected exemplar. The time complexity of the method is analyzed.

Definition 3. Let $IS = (U,A,V,f)$ be a categorical data table and $x \in U$ be an exemplar of the lth cluster C_l. CA_x to be a set of the candidate cluster centers of C_l is defined as

$$CA_x = \{ca_1, \ldots, ca_{|A|}\},$$

where $ca_i \in R$ is a mode of the set $Q_i = \{y \in U|d(y,x) \leqslant i\}$, $1 \leqslant i \leqslant |A|$.

In Table 3, the method for computing CA_x is described where x is given. Table 3 shows that the time complexity of computing CA_x is $O(|U||V|)$.

According to Definition 3, we know that if $1 \leqslant i < j \leqslant m$, $Q_i \subseteq Q_j$, which means that as the value of i increases, the number of the neighbor objects of the exemplar x contained by Q_i increases. In other words, ca_j contains more information of the neighbor objects than ca_i. However, for the value of i, bigger is not better. Because $Q_i = U$ and ca_i is a mode of U when $i = m$. That tells us that the excessive amount of objects contained by Q_i weakens the representability of ca_i to the exemplar x. Therefore, we should integrate the partial neighbor objects of the exemplar. It is important to choose an appropriate ca_i from CA_x.

Next, we define a criterion to select a point from the candidates as the first cluster center. From the candidates, we select a point which is with the higher density and farther from the mode z of U but closer to the exemplar. We give the explanations about the criterion that (1) the selected point with the higher density means that there are more objects around it; (2) the selected point farther from the mode z of U reduces the possibility to be boundary point among clusters; (3) the selected point closer to the exemplar means that it better represents the exemplar x in the cluster.

Definition 4. Let $IS = (U,A,V,f)$ be a categorical data table, z be a mode of U,x be an exemplar of the first cluster C_1 and CA_x be a set of the candidate cluster centers of C_1. For any $y \in CA_x$, the possibility of y to be a cluster center of C_1 is defined as

$$Pos\_center_{C_1}(y) = Dens(y) + d(y,z) - d(y,x).$$

The selection criterion of the rest cluster centers is different from the first cluster center. For selection of the rest cluster centers, we consider the distances between points and other selected cluster centers, instead of the distances between points and the mode z of U. The larger the distances between the point and other selected cluster centers are, the more distinct the point is from other selected cluster centers. In the criterion, we do not consider the distance between the point and the mode z of U, because we can not ignore the fact that z also may be a cluster center in some situations, for example, the data set is imbalanced.

Definition 5. Let $IS = (U,A,V,f)$ be a categorical data table and $Z_l = \{z_1, z_2, \ldots, z_l\}$ be a set of the obtained cluster centers, where $0 < l < k$. For any $x \in U$, the possibility of x to be an exemplar of the $l + 1$ cluster C_{l+1} is defined as

$$Pos\_Exemplar_{C_{l+1}}(x) = Dens(x) + \min_{i=1}^{l} d(x, z_i).$$

From the candidates, we select a point as the cluster center which is with the higher density and farther from other selected cluster centers but closer to the exemplar.

Definition 6. Let $IS = (U, A, V, f)$ be a categorical data table, $Z_l = \{z_1, z_2, \ldots, z_l\}$ be a set of the chosen cluster centers, where $l < k$, x be an exemplar of the $l + 1$th cluster C_{l+1} and CA_x be a set of the candidate cluster centers of C_l. For any $y \in CA_x$, the possibility of y to be a cluster center of C_{l+1} is defined as

$$Pos\_center_{C_{l+1}}(y) = Dens(y) + \min_{i=1}^{l} d(y, z_i) - d(y, x).$$

In the following, a new initialization method for categorical data is described in Table 4.

The time complexity of the proposed initialization method is composed of two parts. First, we obtain a mode of U and calculate densities of all objects in U, whose time complexity is $O(|U||V|)$. Second, the computing of the initial cluster centers will take $O(|U|k^2 + |U||V|k + |A|k^2)$ steps. Therefore, the whole time complexity of the proposed method is $O(|U||V| + |U|k^2 + |U||V|k + |A|k^2)$. In Table 5, the time complexities of Cao's methods and the proposed method are showed. The comparison illustrates that the proposed method requires slightly more computational times than Cao's method. It is an expected outcome since the construction and selection of candidates requires some additional arithmetic operations. However, according to the analysis of the time complexity, the computational complexity of the proposed method is still scalable which is linear with respect to the number of data objects, i.e., it can cluster large categorical data efficiently.

Table 4
A new initialization method for categorical data.

```
Input: $IS = (U,A,V,f)$ and $k$, where $k$ is the number of cluster desired.
Output: $Centers$.

Let $fr_{a,q}$ denotes the number of categorical objects in $U$ which have the value $q$
    of the attribute $a$, $q \in V_a$, $a \in A$.

Begin
  $Centers = \emptyset$;
  $fr_{i,a,q} = 0$, $1 \leqslant i \leqslant |A|$, $q \in V_a$, $a \in A$;
  For each $x$ in $U$
    For each $a$ in $A$
      For each $q$ in $V_a$
        if $f(y,a) == q$
          $fr_{a,q} = fr_{a,q} + 1$;
        end
      end
    end
  end
  Compute a mode $z$ of $U$;
  For each $x$ in $U$
    Compute $Dens(x)$;
  end
  For $i = 1$ to $k$
    $Pos\_Exemplar_{C_i}(x_{C_i}) = max_{y \in U}\{Pos\_Exemplar_{C_i}(y)\}$;
    /*find the most probable examplar $x_{C_i}$ of the $i$th cluster center*/
    Obtain $CA_{x_{C_i}} = \{ca_1, \ldots, ca_m\}$ using the emthod in Table 3;
    For $j = 1$ to $|A|$
      Compute $Dens(ca_j)$;
    end
    $Pos\_center_{C_i}(z_i) = max_{ca_j \in CA_{x_{C_i}}}\{Pro\_center_{C_i}(ca_j)\}$;
    /*Find the $i$th cluster center $z_i$*/
    $Centers = Centers \bigcup \{z_i\}$;
  end
end
```

Table 5
The time complexities of different initialization methods.

Algorithms	Time complexity
Cao's method	$O(\|U\|\|V\| + \|U\|k^2)$
Proposed method	$O(\|U\|\|V\| + \|U\|k^2 + \|U\|\|V\|k + \|A\|k^2)$

Table 6
The initial cluster centers obtained using different initialization methods.

Algorithms	Cluster centers	a_1	a_2	a_3	a_4
Cao's method	z_1	C	E	B	B
	z_2	E	G	C	C
	z_3	C	B	E	E
Proposed method	z_1	B	B	B	B
	z_2	C	C	C	C
	z_3	E	E	E	E

Table 7
Cluster recovery for the data set in Table 2 with different initial cluster centers.

Clusters found	Objects in clusters	Cao's method			Proposed method		
		D_1	D_2	D_3	D_1	D_2	D_3
C_1	4	2	0	2	4	0	0
C_2	4	0	4	0	0	4	0
C_3	4	2	0	2	0	0	4

We use the above example in Table 2 to compare it with Cao's method and demonstrate performance of the proposed initialization method. Table 6 shows the initial cluster centers obtained using Cao's method and the proposed method. We use the k-modes algorithm with the different initial cluster centers to cluster the data set in Table 2. In Table 7, the clustering results are displayed. Tables 6 and 7 illustrate that the proposed method can obtain the better initial cluster centers than Cao's method for clustering the data set in Table 2.

4. Experimental analysis

In this section, in order to evaluate the performance and scalability of the proposed initialization method, some standard data sets are downloaded from the UCI Machine Learning Repository (2010). All missing attribute values are treated as special values. In the performance analysis, we introduce an evaluation method (Yang, 1999) and compare the clustering results of the k-modes algorithm based on different initialization methods including random initialization method, Cao's method and the proposed method. At random initialization method, we carried out 100 runs of the k-modes algorithm on these standard data sets, respectively. In the scalability analysis, we test the proposed algorithm in connect-4 data set from UCI (UCI Machine Learning Repository, 2010).

4.1. Performance analysis

To evaluate the performance of clustering results, an evaluation method is introduced (Yang, 1999). If a data set contains k classes for a given clustering, let a_i denote the number of data objects that are correctly assigned to 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 accuracy, precision and recall are defined as follow: $AC = \frac{\sum_{i=1}^{k} a_i}{|U|}$, $PR = \frac{\sum_{i=1}^{k}\left(\frac{a_i}{a_i+b_i}\right)}{k}$, $RE = \frac{\sum_{i=1}^{k}\left(\frac{a_i}{a_i+c_i}\right)}{k}$, respectively.

We present comparative results of clustering on soybean data, lung cancer data, zoo data, dermatology data, breast cancer data and mushroom data, respectively.

Table 8
Cluster recovery for the soybean data with the initial cluster centers computed by the proposed method.

Clusters found	Objects in cluster	Coming from			
		I	II	III	IV
C_1	10	0	10	0	0
C_2	10	0	0	10	0
C_3	10	10	0	0	0
C_4	47	0	0	0	47

Table 9
Comparison of clustering results of different initialization methods on the soybean data.

The *k*-modes algorithm	Random	Cao's method	Proposed method
AC	0.8564	1.0000	1.0000
PR	0.9000	1.0000	1.0000
RE	0.8402	1.0000	1.0000

Table 10
Cluster recovery for the lung cancer data with the initial cluster centers computed by the proposed method.

Clusters found	Objects in cluster	Coming from		
		I	II	III
C_1	12	7	4	1
C_2	6	0	0	6
C_3	14	2	9	3

Table 11
Comparison of clustering results of different initialization methods on the lung cancer data.

The *k*-modes algorithm	Random	Cao's method	Proposed method
AC	0.5363	0.5000	0.6875
PR	0.6033	0.5584	0.7421
RE	0.5396	0.5014	0.6900

Table 12
Cluster recovery for the zoo data with the initial cluster centers computed by the proposed method.

Clusters found	Objects in cluster	Coming from						
		I	II	III	IV	V	VI	VII
C_1	22	22	0	0	0	0	0	0
C_2	11	0	0	0	0	0	8	3
C_3	16	0	0	3	13	0	0	0
C_4	20	19	0	1	0	0	0	0
C_5	20	0	20	0	0	0	0	0
C_6	5	0	0	1	0	4	0	0
C_7	7	0	0	0	0	0	0	7

Table 13
Comparison of clustering results of different initialization methods on the zoo data.

The *k*-modes algorithm	Random	Cao's method	Proposed method
AC	0.8356	0.8812	0.9208
PR	0.8186	0.8702	0.8985
RE	0.6123	0.6714	0.8143

4.1.1. Soybean data

The soybean data set has 47 records, each of which is described by 35 attributes. Each record is labeled as one of the four diseases: Diaporthe Stem Canker, Charcoal Rot, Rhizoctonia Root Rot, and Phytophthora Rot. Except for Phytophthora Rot which has 17 records, all other diseases have 10 records each. The cluster recovery result of the *k*-modes algorithm with the proposed method on the soybean data is summarized in Table 8. The comparison of clustering results of different initialization methods on the soybean data is presented in Table 9.

Table 14
Cluster recovery for the dermatology data with the initial cluster centers computed by the proposed method.

Clusters found	Objects in cluster	Coming from					
		I	II	III	IV	V	VI
C_1	43	43	0	0	0	0	0
C_2	70	0	0	70	0	0	0
C_3	18	0	0	0	0	0	18
C_4	73	15	7	0	4	45	2
C_5	108	0	54	2	45	7	0
C_6	54	54	0	0	0	0	0

Table 15
Comparison of clustering results of different initialization methods on the dermatology data.

The *k*-modes algorithm	Random	Cao's method	Proposed method
AC	0.6870	0.7486	0.7760
PR	0.7633	0.8801	0.8527
RE	0.5751	0.6091	0.7482

Table 16
Cluster recovery for the breast cancer data with the initial cluster centers computed by the proposed method.

Clusters found	Objects in cluster	Coming from	
		I	II
C_1	229	15	214
C_2	470	443	27

Table 17
Comparison of clustering results of different initialization methods on the breast-cancer data.

The *k*-modes algorithm	Random	Cao's method	Proposed method
AC	0.8461	0.9113	0.9399
PR	0.8700	0.9292	0.9385
RE	0.7833	0.8773	0.9276

Table 18
Cluster recovery for the mushroom data with the initial cluster centers computed by the proposed method.

Clusters found	Objects in cluster	Coming from	
		I	II
C_1	3164	70	3094
C_2	4960	4138	822

Table 19
Comparison of clustering results of different initialization methods.

The *k*-modes algorithm	Random	Cao's method	Proposed method
AC	0.7318	0.8754	0.8902
PR	0.7520	0.9019	0.9061
RE	0.7278	0.8709	0.8867

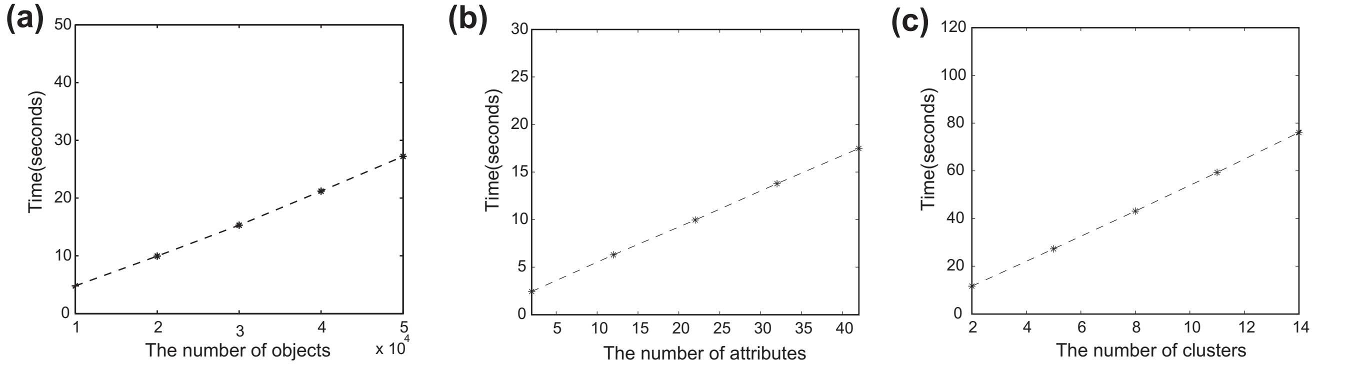

Fig. 1. (a) Computational times for different numbers of objects. (b) Computational times for different numbers of attributes. (c) Computational times for different numbers of clusters.

4.1.2. Lung cancer data

Lung cancer data set contains 32 instances described by 56 categorical attributes. The data set has three classes. The cluster recovery result of the *k*-modes algorithm with the proposed method on the lung cancer data is summarized in Table 10. The comparison of clustering results of different initialization methods on the lung cancer data is presented in Table 11.

4.1.3. Zoo data

Zoo data set contains 101 elements described by 17 Boolean-valued attributes and 1 type attribute. Data set with 101 Elements belong to seven classes. The cluster recovery result of the *k*-modes algorithm with the proposed method on the zoo data is summarized in Table 12. The comparison of clustering results of different initialization methods on the zoo data is presented in Table 13.

4.1.4. Dermatology data

Dermatology data set contains 366 elements and 33 categorical attributes. It has six clusters: psoriasis (112 data objects), seboreic dermatitis (61 data objects), lichen planus (72 data objects), pityriasis rosea (49 data objects), cronic dermatitis (52 data objects) and pityriasis rubra pilaris (20 data objects). The cluster recovery result of the *k*-modes algorithm with the proposed method on the dermatology data is summarized in Table 14. The comparison of clustering results of different initialization methods on the dermatology data is presented in Table 15.

4.1.5. Breast cancer data

Breast cancer data set consists of 699 data objects and 9 categorical attributes. It has two clusters Benign (458 data objects), Malignant (241 data objects). The cluster recovery result of the *k*-modes algorithm with the proposed method on the breast cancer data is summarized in Table 16. The comparison of clustering results of different initialization methods on the breast cancer data is presented in Table 17.

4.1.6. Mushroom data

Mushroom data set consists of 8124 data objects and 23 categorical attributes. Each object belongs to one of two classes, edible (4208 objects) and poisonous (3916 objects). The cluster recovery result of the *k*-modes algorithm with the proposed method on the mushroom data is summarized in Table 18. The comparison of clustering results of different initialization methods on the mushroom data is presented in Table 19.

From the above experiential results, for the *k*-modes algorithm, we can see that the proposed method is superior to Cao's method and random initialization method with respect to *AC*, *PR*, *RE*, respectively.

4.2. Scalability analysis

To test the scalability of the new algorithm, we choose Connect-4 data set from UCI. The data set contains 67,557 objects and 42 categorical attributes. It has three class: win (44,473), loss (16,635) and draw (6449). The computational results are performed by using a machine with an Intel Q9400 and 2G RAM. The computational times of the proposed algorithm are plotted with respect to the number of objects, attributes and clusters, while the other corresponding parameters are fixed.

Fig. 1a shows the computational times against the number of objects, while the number of attributes is 42 and the number of clusters is 3. Fig. 1b shows the computational times against the number of attributes, while the number of clusters is 3 and the number of objects is 30,000. Fig. 1c shows the computational times against the number of clusters, while the number of attributes is 42 and the number of objects is 30,000. According to the figures, we can see that the proposed method is scalable, i.e., it can get the initial cluster centers of categorical data efficiently.

5. Conclusions

Categorical data are ubiquitous in real-world databases. The development of the *k*-modes algorithm was motivated to solve this problem. However, the clustering algorithm need to rerun many times with different initializations in an attempt to find a good solution. Moreover, this 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. In this paper, a new initialization method for categorical data clustering has been proposed by taking into account the distance between the objects and the density of the object and overcomes shortcomings of the existing initialization methods. Furthermore, the time complexity of the proposed method has been analyzed. We tested the proposed method using seven real world data sets from UCI Machine Learning Repository and experimental results have shown that the proposed method is superior to other initialization methods in the *k*-modes algorithm.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Nos. 71031006 and 70971080), GRF: CityU 112809 of Hong Kong SAR Government, the Foundation of Doctoral Program Research of Ministry of Education of China (No. 20101401110002), the National Key Basic Research and Development Program of China (973) (No. 2011CB311805), and the Technology Research Development Projects of Shanxi (No. 2007103).

References

Aggarwal, C. C., Magdalena, C., & Yu, P. S. (2002). Finding localized associations in market basket data. *IEEE Transactions on Knowledge and Data Engineering, 14*(1), 51–62.

Anderberg, M. R. (1973). *Cluster analysis for applications*. Academic.

Andreopoulos, B., An, A., & Wang, X. (2005). Clustering the internet topology at multiple layers. *WSEAS Transactions on Information Science and Applications, 2*, 1625–1634.

Arthur, D., & Vassilvitskii, S. (2007). *K*-means++: The advantages of careful seeding. In *Proceedings 18th annual ACM-SIAM symposium on discrete algorithms (SODA'07)* (pp. 1027–1035).

Babu, G. P., & Murty, M. N. (1993). A near-optimal initial seed value selection for *k*-means algorithm using genetic algorithm. *Pattern Recognition Letters, 14*, 763–769.

Ball, G. H., & Hall, D. J. (1967). A clustering technique for summarizing multivariate data. *Behavioral Science, 12*, 153–155.

Barbara, D., Couto, J., & Li, Y. (2002). COOLCAT: An entropy-based algorithm for categorical clustering. In *Proceedings of the eleventh international conference on information and knowledge management* (pp. 582–589).

Barbara, D., & Jajodia, S. (Eds.). (2002). *Applications of data mining in computer security*. Dordrecht: Kluwer.

Baxevanis, A., & Ouellette, F. (2001). *Bioinformatics: A practical guide to the analysis of genes and proteins* (2nd ed.). NY: Wiley.

Bradley, P. S., & Fayyad, U. M. (1998). Refining initial points for *k*-means clustering. In J. Sharlik (Ed.), *Proceedings of 15th international conference on machine learning (ICML98)* (pp. 91–99). San Francisco, CA: Morgan Kaufmann.

Bradley, P. S., Mangasarian, O. L., & Street, W. N. (1997). Clustering via concave minimization. In M. C. Mozer, M. I. Jordan, & T. Petsche (Eds.). *Advances in neural information processing system* (Vol. 9, pp. 368–374). MIT Press.

Brendan, J. F., & Delbert, D. (2007). Clustering by passing messages between data points. *Science, 15*(16), 972–976.

Cao, F. Y., Liang, J. Y., & Bai, L. (2009). A new initialization method for categorical data clustering. *Expert Systems with Applications, 33*(7), 10223–10228.

Ester, M., Kriegel, H., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Evangelos Simoudis, Jiawei Han, & Usama M. Fayyad (Eds.), *Proceedings of the second international conference on knowledge discovery and data mining (KDD-96)* (pp. 226–231). AAAI Press.

Gowda, K. C., & Diday, E. (1999). Symbolic clustering using a new dissimilarity measure. *Pattern Recognition, 24*(6), 567–578.

Huang, Z.X. (1997). A fast clustering algorithm to cluster very large categorical data sets in data mining. In *Proceeding SIGMOD workshop research issues on data mining and knowledge discovery* (pp. 1–8).

Huang, Z. X. (1998). Extensions to the *k*-means algorithm for clustering large data sets with categorical values. *Data Mining Knowledge Discovery, 2*(3), 283–304.

Huang, Z. X., Ng, M., Rong, H., & Li, Z. (2005). Automated variable weighting in *k*-means type clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*(5), 657–668.

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data*. Prentice Hall.

Khan, S. S., & Ahmad, A. (2004). Cluster center initialization algorithm for *k*-means clustering. *Patter Recognition Letters, 25*, 1293–1302.

Krishna, K., & Murty, M. N. (1999). Genetic *k*-means algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, 29*(3), 433–439.

Laszlo, M., & Mukherjee, S. (2006). A Genetic algorithm using hyper-quadtrees for low-dimensional *k*-means clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28*(4), 533–543.

Laszlo, M., & Mukherjee, S. (2007). A genetic algorithm that exchanges neighboring centers for *k*-means clustering. *Pattern Recognition Letters, 28*(16), 2359–2366.

Liang, J. Y., & Li, D. Y. (2005). *Uncertainty and knowledge acquisition in information systems*. Beijing, China: Science Press.

Liang, J. Y., Wang, J. H., & Qian, Y. H. (2009). A new measure of uncertainty based on knowledge granulation for rough sets. *Information Sciences, 179*(4), 458–470.

MacQueen, J.B. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of fifth symposium on mathematical statistics and probability* (Vol. 1, pp. 281–297).

Pawlak, Z. (1991). *Rough sets-theoretical aspects of reasoning about data*. Dordrecht, Boston, London: Kluwer Academic Publishers.

Pen, J. M., Lozano, J. A., & Larraaga, P. (1999). An empirical comparsion of four initalization methods for the *k*-means algorithm. *Pattern Recognition Letter, 20*, 1027–1040.

Qian, Y. H., Liang, J. Y., Pedrycz, W., & Dang, C. Y. (2010). Positive approximation: An accelerator for attribute reduction in rough set theory. *Artificial Intelligence, 174*(5–6), 597–618.

Sun, Y., Zhu, Q. M., & Chen, Z. X. (2002). An iterative initial-points refinement algorithm for categorical data clustering. *Pattern Recognition Letters, 23*, 875–884.

UCI Machine Learning Repository (2010). <http://www.ics.uci.edu/mlearn/MLRepository.html>.

Wang, J., & Karypis, G. (2006). On efficiently summarizing categorical databases. *Knowledge and Information Systems, 9*(1), 19–37.

Wrigley, N. (1985). *Categorical data analysis for geographers and environmental scientists*. London: Longman.

Wu, S., Jiang, Q. S., & Huang, Z. X. (2007). A new initialization method for categorical data clsutering. *Lecture Notes in Computer Science, 4426*, 972–980.

Yang, Y. M. (1999). An evaluation of statistical approaches to text categorization. *Journal of Information Retrieval, 1*(1–2), 67–88.