Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.12785/amis/090621

A Novel Cluster Center Initialization Method for the k-Prototypes Algorithms using Centrality and Distance

Jinchao Ji[1,2,3,*], *Wei Pang*[4], *Yanlin Zheng* [1,2], *Zhe Wang*[3,5], *Zhiqiang Ma*[1,2,*] *and Libiao Zhang*[1,2,*]

[1] School of Computer Science and Information Technology, Northeast Normal University, Changchun, 130117, China
[2] Key Lab of Intelligent Information Processing of Jilin Universities, Northeast Normal University, Changchun, 130117, China
[3] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, 130012, China
[4] School of Natural and Computing Sciences, University of Aberdeen, Aberdeen, AB24 3UE, UK
[5] College of Computer Science and Technology, Jilin University, Changchun, 130012, China

Received: 8 Sep. 2014, Revised: 30 Mar. 2015, Accepted: 31 Mar. 2015
Published online: 1 Nov. 2015

Abstract: The k-prototypes algorithms are well known for their efficiency to cluster mixed numeric and categorical data. In k-prototypes type algorithms the initial cluster centers are often determined in a random manner. It is acknowledged that the initial placement of cluster centers has a direct impact on the performance of the k-prototypes algorithms. However, most of the existing initialization approaches are designed for the k-means or k-modes algorithms, which can only deal with either pure numeric or categorical data, but not the mixture of both. In this paper, we propose a novel cluster center initialization method for the k-prototypes algorithms to address this issue. In the proposed method, the centrality of data objects is introduced based on the concept of neighbor-set, and then both the centrality and distance are exploited together to determine initial cluster centers. The performance of the proposed method is demonstrated by a series of experiments in comparison with that of traditional random initialization method.

Keywords: clustering, data mining, mixed numeric and categorical data, cluster center initialization

1 Introduction

Clustering analysis, an important task in data mining [1, 2], has been applied in a broad range of fields including information retrieval [3], privacy preserving [4], image analysis [5], text analysis [6], and bioinformatics [7]. Clustering is a process of classifying a set of data objects into clusters such that similar data objects are in the same cluster and dissimilar ones are in different clusters [2,8,9, 10]. Clustering methods in the literature generally fall into two categories: hierarchical and partitional. Hierarchical clustering algorithms organize a collection of data objects into a dendrogram of the nested partitions by utilizing a divisive or agglomerative strategy. Whilst partitional clustering algorithms divide a set of data objects into the given number of clusters by optimizing an objective cost function.

The k-means and k-modes algorithms are efficient partitional clustering algorithms for numeric and categorical data, respectively. To deal with the data objects with both numeric and categorical attributes, Huang integrated the k-means with k-modes to propose the k-prototypes algorithm [11]. Up till now, several extensions of the k-prototypes algorithm have been proposed, such as the methods introduced by Ahmad & Dey [12], and Ji *et al.* [13,14]. These partitional clustering algorithms generally start with a set of initial cluster centers and frequently end up with different clustering outcomes for different groups of initial cluster centers [15]. Therefore, how to determine the initial cluster centers is an important issue for partitional clustering algorithms, because it directly influences the formation of final clusters [16,17]. In the light of the type of data to be processed, cluster center initialization approaches are mainly classified into three categories: numeric data, categorical data, and mixed data initialization approaches.

A great deal of effort has been made to initialize the cluster centers for k-means algorithm [18]. The method developed by Forgy [19] initializes the cluster centers by

* Corresponding author e-mail: jinchao0374@gmail.com, mazq0431@gmail.com, lbzhang@nenu.edu.cn

the following procedure: data objects in a dataset are first assigned to one of the k clusters randomly, and then the centroids of these initial clusters are taken as the initial cluster centers [1,19]. Another method proposed by Jancey [20] is to allocate each cluster center a synthetic data object generated randomly within the given data space [1,20]. MacQueen [18] presented two different approaches: the first method, which is sensitive to the order of data objects [1], takes the first k data objects in the dataset as initial cluster centers; the second one randomly picks up k data objects in the dataset as initial cluster centers. The second method is based on the hypothesis that random selection may pick up good candidates as cluster centers, and this method has become the standard approach for determining the placement of initial cluster centers in the k-means algorithm [21], although outliers may be selected as centers when using this method [1]. Ball and Hall's method (BH) first takes the center of the entire dataset as the first cluster center, and then picks up the data object which is at least T units away from the existing cluster centers as the next cluster center. This process does not terminate until the expected k cluster centers are obtained. Unlike the BH method, the Simple Cluster Seeking (SCS) method [22] picks up the first data object in the dataset as the first cluster center. Maxmin method [23,24] randomly selects a data object as the first cluster center, then the data object with the greatest minimum-distance to the existing cluster centers is taken as the next cluster center. This process repeats until k cluster centers are achieved [1]. Al-Daoud introduced two approaches: Al-Daoud's method 1 (AD1) [25] and Al-Daoud's method 2 (AD2) [26]. AD1 first uniformly divides the data space into a given number of disjoint hyper-cubes, then randomly picks up a fixed number of data objects from each hyper-cube [1] to form the expected cluster centers. In AD2 the data objects are first ranked on the attribute with the maximum variance, and then these data objects are allocated into k groups along the same attribute, and finally the data objects corresponding to each median are employed to determine the initial cluster centers [1,26]. The k-means++ approach [27] combines MacQueen's second method with the Maximin method to initialize the cluster centers. The cluster center initialization algorithm (CCIA) [28] first selects $k' > k$ centers from the centroids calculated by implementing the k-means algorithm [18] on each attribute, and then merges similar centers to form k initial cluster centers. The Redmond and Heneghan's method (RH) [29] adopts the concept of kd-tree to evaluate the density, and then utilizes a modified Maximin method to initialize the cluster centers. Cao *et al.* presented an initialization approach on the basis of neighborhood-based rough set model [30]. Yi *et al.* selected the data objects which belong to high density area as initial cluster centers [31]. Kumar, Chhabra, and Kumar introduced an initialization approach by adopting the biogeography-based optimization [32].

Up till now there exist several approaches to performing the cluster center initialization for the k-modes algorithm. Huang introduced two approaches [8]: the first approach takes the first k distinct data objects as the initial cluster centers; the second one first assigns the most frequent categories to the initial k modes equally, and then the most similar data objects to these modes are picked up as the initial cluster centers [8]. Sun *et al.* suggested an initialization approach which adopted an iterative initial-point refinement procedure devised by Bradley and Fayyad [21] to improve the accuracy and reproducibility of the clustering results [33,34,35]. Khan and Ahmad [36] integrated Hamming distance with density-based multi-scale data condensation technique [37] to determine the initial cluster centers. Two farthest-point heuristics were introduced to initialize cluster centers [38]. The first heuristic takes an arbitrary data object as the first cluster center, and then picks up the data object which has the farthest distance to the nearest cluster centers as the next cluster center. This process does not terminate until the expected k cluster centers are acquired. Unlike the first one, the second heuristic adopts a scoring function to assess the data objects in a dataset, and then takes the data object with the highest score as the first cluster center. Wu *et al.* applied the concept of density to initialize cluster centers [17]. However, the determination of sub-samples in this method introduces the randomness, which may incur unstable and non-repeatable clustering outcomes [33,34]. Cao *et al.* [16] integrated the distance between data objects with the density of the data objects to initialize cluster centers. In the approach proposed by Cao *et al.* [16], a boundary point might be chosen as the initial cluster center [33]. To overcome this deficiency, Bai *et al.* presented an approach on the basis of the cluster exemplar [34]. Khan and Kant proposed an initialization approach by utilizing the concept of evidence accumulation [33,39]. In this method the k-modes algorithm [8] with random initialization was first run N times to get a mode-pool, and then the most diverse set of modes were selected from the available mode-pool as initial cluster centers. Moreover, Khan and Ahmad [33] proposed another method to determine initial cluster centers by adopting multiple attribute clustering.

From the above one can see that the cluster center initialization of both the k-means and k-modes algorithms has been well studied. However, it is acknowledged that data objects with both numeric and categorical attributes are ubiquitous in real-world applications. The concurrence of numeric and categorical attributes makes the initialization methods devised for the k-means or k-modes algorithms unsuitable for the k-prototypes algorithms. To the best of our knowledge, the initial cluster centers are determined in a random manner in the k-prototypes type algorithms. This initialization approach may lead to unstable and non-reproducible clustering outcomes. Thus, it is necessary to develop a more effective initialization method specifically for k-prototypes type algorithms.

In this paper, we propose a novel cluster center initialization method for k-prototypes algorithms. In our method, the centrality of data object is measured on the basis of the neighbor-set model, and then the initial cluster centers are determined by integrating the centrality of a data object with the distance between data objects. The proposed initialization method is used along with the k-prototypes algorithm proposed by Huang [11], one of the most well-known k-prototypes clustering algorithms. The time and space complexity of our proposed approach is analyzed, and the comparison with traditional methods demonstrates the effectiveness of our approach.

The rest of this paper is organized as follows: we first review the k-prototypes algorithm proposed by Huang in Section 2. This is followed by the presentation of our initialization method in Section 3. In Section 4, we report the experimental results, which demonstrate the advantages of the proposed method. Finally, we draw conclusions and explore future work in Section 5.

2 The k-prototypes algorithm

In this section, we first introduce the notations used for representing mixed data: let $X = \{x_1, x_2, \cdots, x_n\}$ denote a dataset consisting of n data objects and $x_i (1 \leq i \leq n)$ be a data object characterized by m attributes $A_1, A_2, \cdots, A_m$. Each attribute A_j has a domain of values denoted by $Dom(A_j)$. The domain of attribute related to mixed data has two types: numeric and categorical. The numeric domain is a set of real numbers, whereas the categorical domain is usually represented by $Dom(A_j) = \{a_j^1, a_j^2, \cdots, a_j^t\}$, where t is the number of categorical values for the categorical attribute A_j. A data object x_i is generally represented as a m-dimension vector $[x_{i1}^r, x_{i2}^r, \cdots, x_{ip}^r, x_{ip+1}^c, x_{ip+2}^c, \cdots, x_{im}^c]$, where the first p items with the superscript r are numeric values and the rest $(m-p)$ are categorical ones.

The k-prototypes algorithm was first developed by Huang in [11]. The aim of the k-prototypes algorithm is to divide the dataset X into k clusters by minimizing the following cost function:

$$E(U,Q) = \sum_{l=1}^{k} \sum_{i=1}^{n} u_{il} dis(x_i, Q_l). \quad (1)$$

Here Q_l is the center or prototype of the cluster l; $u_{il} (0 \leq u_{il} \leq 1)$ is an element of the partition matrix $U_{n \times k}$; and $dis(x_i, Q_l)$ is the distance measure given by:

$$dis(x_i, Q_l) = \sum_{j=1}^{p} (x_{ij}^r - q_{lj}^r) + \sum_{j=p+1}^{m} \mu_l \alpha(x_{ij}^c, q_{lj}^c). \quad (2)$$

In Equation (2), $\alpha(x_{ij}^c, q_{lj}^c)$ is defined as follows:

$$\alpha(x_{ij}^c, q_{lj}^c) = \begin{cases} 0 & \text{if } x_{ij}^c = q_{lj}^c, \\ 1 & \text{if } x_{ij}^c \neq q_{lj}^c, \end{cases} \quad (3)$$

and μ_l is a weight for categorical attributes in a cluster l. In Equation (2), q_{lj}^r is the mean of the jth numeric attribute in a cluster l, and q_{lj}^c is the mode of the jth categorical attribute in a cluster l. The process of Huang's k-prototypes algorithm is given as follows:

Step 1. Randomly pick up k data objects from the dataset X as the initial prototypes of clusters.

Step 2. For each data object in X, assign it to the cluster whose prototype is nearest to this data object in terms of Equation (2). After each assignment, update the prototype of relevant clusters.

Step 3. Re-evaluate the dissimilarity between data objects and the current prototypes after all data objects have been assigned to clusters. If a data object is found that its nearest prototype belongs to another cluster rather than the current one, reassign this data object to that cluster and update the prototypes of both clusters.

Step 4. After a full circle test of X, if no data objects have changed clusters, terminate the algorithm; otherwise return to *Step 3*.

3 Our proposed algorithm

In this section, we propose a new cluster center initialization method based on centrality and distance (CCICD), for k-prototypes types algorithms. It is well known that the initial cluster centers should be separated since clusters are separated in the attribute space. Moreover, the location of a cluster center should be in the central region of the cluster rather than the periphery of cluster.

According to this idea, we first introduce the concept of neighbor-set, which is the set of neighbors, to measure the centrality of data object, and then integrate the centrality measure with distance to evaluate the possibility of a data object to be an initial cluster center.

Definition 1. Let $X = \{x_1, x_2, \cdots, x_n\}$ be the dataset with m mixed numeric and categorical attributes, for any data object $x_i \in X$ the neighbor-set of x_i, denoted by $NborS(x_i)$, is given by:

$$NborS(x_i) = \{x_t \mid dis(x_i, x_t) \leq \sigma, \text{ for } x_t \in X\} \quad (4)$$

where $\sigma > 0$ is the neighbor threshold which is set in advance; the higher the value of σ is, the more data objects the neighbor-set $NborS(x_i)$ includes; $dis(\cdot)$ is the distance measure designed for mixed data. To make the contribution of the numeric attributes and categorical attributes on the same scale, the distance measure $dis(\cdot)$ in our initialization method CCICD is given as follow:

$$dis(x_i, x_j) = \sum_{l=1}^{p} \left(\frac{x_{il}^r - x_{jl}^r}{\lambda_l} \right)^2 + \sum_{l=p+1}^{m} \alpha(x_{il}^c, x_{jl}^c) \quad (5)$$

where $\lambda_l = max_l - min_l$ is the normalization factor for numeric attribute l; max_l and min_l is the maximum and

minimum value for attribute l in the dataset X, respectively. According to Equation (5), the distance $dis(\cdot)$ between any two data objects falls into the range from 0 to m, i.e., $0 \leq dis \leq m$, and the neighbor threshold therefore is between 0 and m as well, i.e., $0 < \sigma \leq m$.

Definition 2. Let $X = \{x_1, x_2, \cdots, x_n\}$ be the dataset with m mixed numeric and categorical attributes, and $NborS(x_i)$ be the neighbor-set of data object x_i. For any data object $x_i \in X$, the centrality of data object x_i is defined as follows:

$$Cen(x_i) = \frac{|\, NborS(x_i)\,|}{|\, NborSMax\,|} \tag{6}$$

where $|\cdot|$ is the cardinality of a set, and $NborSMax = \max_{x_{i \in X}}(NborS(x_i))$ is the neighbor-set which has the most elements. $Cen(x_i)$ is used to measure the centrality of the data object in a cluster. The larger the value of $Cen(x_i)$ is, the more central this data object situates in the cluster. Based on the concept of centrality, the probability of a data object to be the first cluster center is given by the following definition:

Definition 3. Let $X = \{x_1, x_2, \cdots, x_n\}$ be the dataset with m mixed numeric and categorical attributes. For any data object $x_i \in X$, the probability of data object x_i in the dataset X to be the first cluster center is given as:

$$Pro_1(x_i) = Cen(x_i) \tag{7}$$

Assume there are k cluster centers, and we pick up the data object with the highest centrality as the first cluster center according to Equation (7). However, if the centrality is only considered the data objects in the same cluster might be picked up as initial cluster centers; if the distance is the only factor considered, outliers might be chosen as initial cluster centers. To overcome these issues, we combine the centrality with distance to access the probability of data objects to be the rest of cluster centers as follows.

Definition 4. Let $X = \{x_1, x_2, \cdots, x_n\}$ be the dataset with m mixed numeric and categorical attributes, and $Q_l = \{q_1, q_2, \cdots, q_l\}$ be the set of acquired cluster centers, where $1 \leq l < k$. The probability of data objects in the dataset X to be the $l+1$th cluster center is given by:

$$Pro_{l+1}(x_i) = \min_{q_t \in Q_l} dis(x_i, q_t) \times Cen(x_i) \tag{8}$$

Having introduced the aforementioned four definitions in the context of mixed numeric and categorical data, the process of initializing cluster centers is described as follows:

Input: A mixed dataset X, the desired number of clusters k, neighbor threshold σ.

Output: Cluster centers Q.

Step 1. For each data object x in the dataset X, calculate its probability $Pro_1(x)$ according to Equation 7. Then pick up the data object x with the highest probability value as the first cluster center q_1. Thus $Q_s = q_1$, set $s = 1$.

Step 2. If $s < k$, go to *Step* 3; otherwise output initial cluster centers Q_s, and algorithm terminates.

Step 3. For each data object x in X, evaluate its probability $Pro_{s+1}(x)$ according to Equation (8). Then select the data object x with the highest probability value as the $s+1$th cluster center q_{s+1}. Then $Q_{s+1} = Q_s \bigcup q_{s+1}$, set $s = s+1$, and go to *Step 2*.

The time complexity of the proposed method mainly consists of two parts: the computation of the centrality for each data object in X, and the probability of data object to be the initial cluster center. The computational cost of these two parts are $O(n(n-1)m)$, and $O(n + nmk(k-1))$, respectively. Here n is the number of data objects in the dataset X; m is the number of attributes; and k is the number of clusters. Therefore, the overall time complexity is $O(n(n-1)m + n + nmk(k-1))$. For space complexity, it requires $O(mn)$ to store the dataset X, $O(n)$ to store the centrality matrix of data objects, and $O(km)$ to store cluster centers. Thus, the overall space complexity of our initialization method is $O(mn + n + km)$. The time complexity, and space complexity for random initialization are $O(k)$, and $O(nm + km)$, respectively. For Forgys method [19], the time complexity and space complexity are $O(n + nkm)$, and $O(nm + mT + km)$, respectively. Here, T is the maximum number of different categorical values in all categorical attributes. The time complexity and space complexity for Maxmin method are $O(nmk(k-1))$ and $O(nm + n + km)$, respectively.

4 Experimental results

In this section, to evaluate the performance of our proposed initialization method CCICD (Cluster Center Initialization based on Centrality and Distance), we utilize CCICD along with the k-prototypes algorithm [11] to cluster three real-world mixed datasets: Zoo, Heart Disease and Credit Approval, which are derived from UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets.html). In this research, we adopt two commonly used measures, i.e., the clustering accuracy (AC) [40] and the Rand Index (RI) [41], to assess the quality of clustering results. The clustering accuracy (AC) is given by

$$AC = \frac{\sum_{i=1}^{k} a_i}{n}, \tag{9}$$

where a_i is the number of data objects appearing both in the ith cluster and its corresponding true class, and n is the number of data objects in the dataset. Given a dataset $X = \{x_1, x_2, \cdots, x_n\}$ as well as two partitions of this dataset: $Y = \{y_1, y_2, \cdots, y_{t_1}\}$ and $Y' = \left\{y_1', y_2', \cdots, y_{t_1}'\right\}$, the Rand Index (RI) [41] is given by

$$RI = \frac{\sum_{i=1, j=2; i<j}^{n} \eta_{ij}}{\binom{n}{2}} \tag{10}$$

where

$$\eta_{ij} = \begin{cases} 1 & \text{if there exist } t \text{ and } t' \text{ such that} \\ & \text{both } x_i \text{ and } x_j \text{ are in both } y_t \\ & \text{and } y'_{t'}, \\ 1 & \text{if there exist } t \text{ and } t' \text{ such that} \\ & x_i \text{ in both } y_t \text{ and } y'_{t'} \text{ while } x_j \text{ is} \\ & \text{in neither } y_t \text{ or } y'_{t'}, \\ 0 & \text{otherwise.} \end{cases} \quad (11)$$

The *RI* is calculated by using the true clustering and the clustering obtained from a clustering algorithm. According to these measures, the higher values of *AC* and *RI* indicate a better clustering result. In the performance analysis, we compare the clustering results of the k-prototypes algorithm based on different initialization methods, including the traditional random initialization method, the Forgy method, Maxmin method, and our proposed method CCICD. To make the contribution of the numeric attributes and categorical attributes on the same scale, Equation (5) is adopted in k-prototypes algorithm to evaluate the dissimilarity measure between data objects and cluster centers.

In all experiments, the k-prototypes algorithm based on different initialization methods, i.e., our proposed method, the traditional random initialization method, the Forgy method, and the Maxmin method, is run 20 trials. All algorithms are implemented in Java language and executed on an Intel(R) Core(TM) i7, 3.4GHz, 8GB RAM computer. For each dataset, the parameter k of the k-prototypes algorithm is determined according to the class information, which is not utilized in clustering process. To assess the impact of neighbor threshold σ on our proposed initialization approach CCICD, the AC and RI of k-prototypes algorithm based on CCICD with different neighbor threshold σ is compared. Specifically, we perform our approach CCICD with the neighbor threshold σ varied from 1 to m in increment of 1 since the distance between any two data objects falls into the range from 1 to m. For assessing the performance of initialization methods, the average (Avg.), and the standard deviation (Std.) of performance measures (i.e., AC, and RI) is calculated. A higher value of the average value as well as a lower standard deviation of the performance measures means a better quality of the initialization method.

Zoo dataset contains 101 data objects, each of which has one numeric attribute and 16 categorical attributes. According to the class attribute, the data objects belong to one of the seven classes. Fig. 1 illustrates the impact of neighbor threshold σ on the clustering accuracy (AC) and the Rand Index (RI) of the k-prototypes algorithm using our proposed initialization approach CCICD when clustering the Zoo dataset. From this figure, we can see that the neighbor threshold σ has an important impact on AC and RI, and the maximum is achieved when σ is taken the value from 12 to 16 for AC and from 11 to 16 for RI. Table 1 lists the comparison of clustering results of the random initialization method, the Forgy method, the Maxmin method, and our proposed method CCICD $(\sigma = 12)$ on Zoo dataset according to AC. Table 2 summarizes the comparison of clustering results of the random initialization method, the Forgy method, the Maxmin method, and our proposed method CCICD $(\sigma = 12)$ on Zoo dataset according to RI. From these tables, we can see that the proposed CCICD achieved a bigger average value and lower standard deviation of AC and RI than the other three methods. Therefore, our proposed method outperforms the other three initialization methods on Zoo dataset according to both AC and RI.

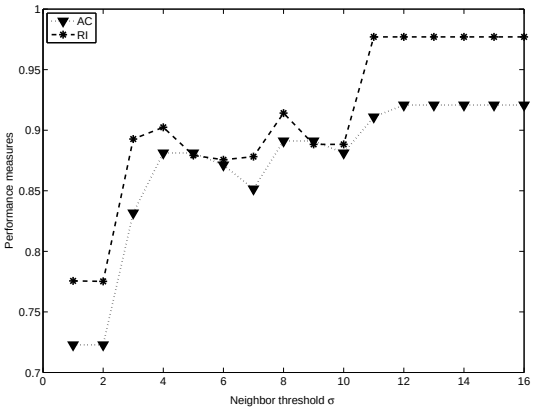


Fig. 1: The impact of neighbor threshold σ on the clustering results of the k-prototypes algorithm with the initialization method CCICD for Zoo dataset

Table 1: The AC of initialization methods on Zoo dataset

Initialization algorithms	AC	
	Ave.	Std.
Random	0.8490	0.0444
Forgy	0.8748	0.0217
Maxmin	0.8906	0.0149
CCICD $(\sigma = 12)$	0.9208	0.0000

Table 2: The RI of initialization methods on Zoo dataset

Initialization algorithms	AC	
	Ave.	Std.
Random	0.8890	0.0458
Forgy	0.9318	0.0256
Maxmin	0.9355	0.0286
CCICD $(\sigma = 12)$	0.9770	0.0000

Heart Disease dataset consists of 303 patient instances, each of which has six numeric attributes and nine categorical attributes. The last two attributes are class attributes. When we take the 15th attribute as its

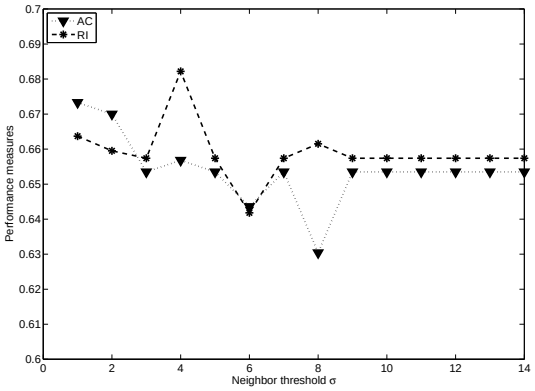


Fig. 2: The impact of neighbor threshold σ on the clustering results of the k-prototypes algorithm with the initialization method CCICD for Heart Disease dataset (first case)

Table 3: The AC of initialization methods on Heart Disease dataset (first case)

Initialization algorithms	AC	
	Ave.	Std.
Random	0.6520	0.0130
Forgy	0.6488	0.0141
Maxmin	0.6493	0.0097
CCICD $(\sigma = 1)$	0.6733	0.0000

Table 4: The RI of initialization methods on Heart Disease dataset (first case)

Initialization algorithms	AC	
	Ave.	Std.
Random	0.6596	0.0182
Forgy	0.6417	0.0175
Maxmin	0.6566	0.0145
CCICD $(\sigma = 4)$	0.6822	0.0000

class attribute, the data objects belong to one of the five classes (s1, s2, s3, s4, and H), and the dataset are characterized by 14 attributes; whereas when we take the 14th attribute as its class attribute, the data objects belong to one of the two classes (buff, sick), and the dataset are characterized by 13 attributes. For the first case, Fig. 2 displays the impact of neighbor threshold σ on the clustering accuracy (AC) and Rand Index (RI) of the k-prototypes algorithm adopting our proposed initialization approach CCICD. From Fig. 2, we can see that the neighbor threshold σ has a significant impact on AC and RI, and the maximum performance are achieved when the neighbor threshold σ equals to 1 for AC and 4 for RI, respectively. Table 3 summarizes the comparison of clustering results of the random initialization method, the Forgy method, the Maxmin method, and our proposed method CCICD $(\sigma = 1)$ on Heart Disease dataset (first case) according to AC. Table 4 lists the comparison of clustering results of the random initialization method, the Forgy method, the Maxmin method, and our proposed method CCICD $(\sigma = 4)$ on Heart Disease dataset (first case) according to RI. From Tables 3 and 4, we can see that the proposed CCICD achieved a bigger average value and lower standard deviation of AC and RI than the other three methods. Therefore, our proposed method outperforms the other three initialization methods on Heart Disease dataset (first case) according to AC and RI, respectively.

For the second case where data objects in Heart Disease dataset are described by 13 attributes, we take the 14th attribute of data object as its class attribute. Fig. 3 illustrates the impact of neighbor threshold σ on the clustering accuracy (AC) and Rand Index (RI) of the k-prototypes algorithm using our proposed initialization approach CCICD. From this figure, we can see that the neighbor threshold σ has no obvious impact on AC and RI for this case. Table 5 lists the comparison of clustering results of the random initialization method, the Forgy method, the Maxmin method, and our proposed method CCICD $(\sigma = 1)$ on Heart Disease dataset (second case) according to AC. Table 6 summarizes the comparison of clustering results of the random initialization method, the Forgy method, the Maxmin method, and our proposed method CCICD $(\sigma = 1)$ on Heart Disease dataset (second case) according to RI. From Tables 5 and 6, we can see that all four initialization methods obtained similar value of AC and RI.

Credit approval dataset contains 690 customer instances from credit card organizations, each of which is described by ten categorical attributes and six numeric attributes. According to the class attribute, the data objects belong to one of the two classes: negative and positive. Fig. 4 shows the impact of neighbor threshold σ on the clustering accuracy (AC) and Rand Index (RI) of the k-prototypes algorithm using our proposed initialization approach CCICD for the Credit dataset. From Fig. 4, we can see that the neighbor threshold σ has a significant impact on the AC and RI, and the maximum of them is achieved when the neighbor threshold σ equals to 1. Table 7 summarizes the comparison of clustering results of the random initialization method, the Forgy method, the Maxmin method, and our proposed method CCICD $(\sigma = 1)$ on Credit dataset according to AC. Table 8 lists the comparison of clustering results of the random initialization method, the Forgy method, the Maxmin method, and our proposed method CCICD $(\sigma = 1)$ on Credit dataset according to RI. From Tables 7 and 8, we can see that the proposed CCICD achieved a bigger average value and lower standard deviation of AC and RI than the other three methods. Therefore, our proposed method outperforms the other three initialization methods on Credit dataset according to AC and RI, respectively.

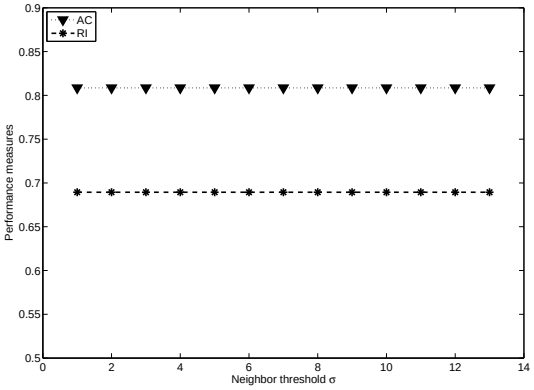


Fig. 3: The impact of neighbor threshold σ on the clustering results of the k-prototypes algorithm with the initialization method CCICD for Heart Disease dataset (second case)

Table 5: The AC of initialization methods on Heart Disease dataset (second case)

Initialization algorithms	AC	
	Ave.	Std.
Random	0.8086	0.0000
Forgy	0.8086	0.0000
Maxmin	0.8086	0.0000
CCICD ($\sigma = 1$)	0.8086	0.0000

Table 6: The RI of initialization methods on Heart Disease dataset (second case)

Initialization algorithms	AC	
	Ave.	Std.
Random	0.6894	0.0000
Forgy	0.6894	0.0000
Maxmin	0.6894	0.0000
CCICD ($\sigma = 1$)	0.6894	0.0000

Table 7: The AC of initialization methods on Credit dataset

Initialization algorithms	AC	
	Ave.	Std.
Random	0.7424	0.0511
Forgy	0.8000	0.0000
Maxmin	0.7672	0.0445
CCICD ($\sigma = 1$)	0.8000	0.0000

Table 8: The RI of initialization methods on Credit dataset

Initialization algorithms	AC	
	Ave.	Std.
Random	0.6219	0.0489
Forgy	0.6795	0.0000
Maxmin	0.6461	0.0429
CCICD ($\sigma = 1$)	0.6795	0.0000

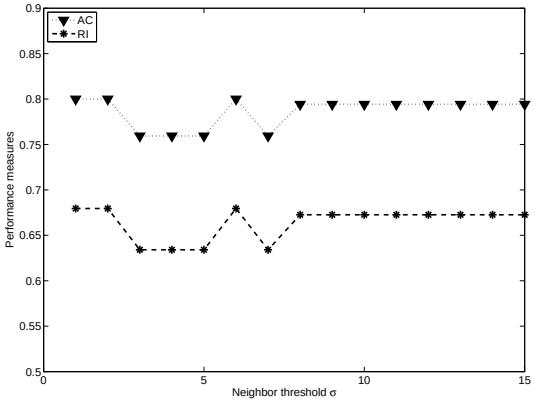


Fig. 4: The impact of neighbor threshold σ on the clustering results of the k-prototypes algorithm with the initialization method CCICD for Credit dataset

Table 9: The average running time of the four algorithms on different datasets

Datasets	Average running time (millisecond)			
	Random	Forgy	Maxmin	CCICD
Zoo	0.05	0.4	4.3	4.5
Heart Disease1	0.05	1.15	8.7	77.15
Heart Disease2	0.05	1.05	2.5	73.65
Credit	0.05	2.95	4.5	443.7

In general, from Figs. 1-4 we can see that the neighbor threshold σ has a significant impact on the result of the CCICD which in turn affects the clustering results of the k-prototypes algorithm. For each dataset (including the same dataset with different case), there exists a suitable value of neighbor threshold for CCICD. Due to the space limitation, we will explore the issue of determination of suitable neighbor threshold in our future work. From Tables 1-8, we can see that the clustering accuracy (AC) and Rand Index (RI) of the k-prototypes algorithm is higher and more stable across different runs when adopting our proposed approach with suitable neighbor threshold σ than that of the same algorithm when using the other three initialization methods in most cases. The reason is that the initial cluster centers selected by the proposed CCICD method are at or close to the real cluster centers for each run. Moreover, Table 9 lists the average runtime of all four initialization methods on different datasets. From this table, we can see that the CCICD needs more time than the other three methods. This is consistent with the analysis of the time complexity in Section 3.

5 Conclusions and Future work

In many real-world applications, data objects are often described by both numeric and categorical attributes. Regarding this the k-prototypes algorithms have been developed to perform the clustering tasks on such mixed data, and these algorithms have been proven to be efficient. However, the initialization of cluster centers is an important procedure in the k-prototypes type algorithms and it may have a big impact on the performance of a clustering algorithm: a good initialization method may significantly improve the clustering result. Currently, the research on the initialization of cluster centers for clustering algorithms mostly focuses on the k-means or k-modes algorithms, which can only deal with numeric or categorical data.

In this paper, we proposed a novel initialization approach CCICD to address this issue for the k-prototypes algorithms, and to start with we use Huangs k-prototypes algorithm with the proposed CCICD. In our method, we introduce the concept of neighbor-set to assess the centrality of data object in a cluster, and then integrate the centrality of data object with the distance between data objects to evaluate the probability of a data object to be a cluster center. We adopt the k-prototypes algorithm based on our proposed initialization approach CCICD to test the impact of neighbor threshold σ on CCICD, and the performance of CCICD on three real-world datasets. The experimental results demonstrate that the neighbor threshold σ has a significant impact on the performance of the proposed CCICD, and when used along with Huang's k-prototypes algorithm the CCICD with a suitable neighbor threshold σ outperforms the other three initialization methods according to the clustering accuracy (AC), and the Rand Index (RI) in most cases. Due to space limitation of the paper, at the moment we use Huang's k-prototypes algorithm to test our approach. However, we point out that by utilizing the corresponding distance measure, our CCICD can also be adapted to other k-prototypes type algorithms.

For the near future, we will explore the issue of determining suitable neighbor threshold for the proposed initialization method CCICD, and investigate the acceleration issue of evaluating the centrality of data objects.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant Nos. (21127010, 61202309), China Postdoctoral Science Foundation under Grant No. 2013M530956, Science and Technology Development Plan of Jilin province under Grant No. 20140520068JH, Fundamental Research Funds for the Central Universities under No. 14QNJJ028, the open project program of Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University under Grant No. 93K172014K07, the 2014 Industrial Technology Research and Development Special Project of Jilin Province, the 2015 Department of Education 12th Five-Year Science and Technology Research Planning Projects of Jilin Province.

The authors are grateful to the anonymous referee for a careful checking of the details and for helpful comments that improved this paper.

References

[1] M.E. Celebi, H.A. Kingravi, and P.A. Vela, A comparative study of efficient initialization methods for the k-means clustering algorithm. Expert Systems with Applications, **40**, 200-210 (2013).

[2] A.K. Jain, M.N. Murty, and P.J. Flynn, Data clustering: A review. ACM Computing Surveys, **31**, 264-323 (1999).

[3] G. Bordogna, and G. Pasi, A quality driven hierarchical data divisive soft clustering for information retrieval. Knowledge-Based Systems, **26**, 9-19 (2012).

[4] M.Z. Islam, and L. Brankovic, Privacy preserving data mining: A noise addition framework using a novel clustering technique. Knowledge-Based Systems, **24**, 1214-1223 (2011).

[5] C. Bogner, B.T.Y. Widemann, and H. Lange, Characterising flow patterns in soils by feature extraction and multiple consensus clustering. Ecological Informatics, **15**, 44-52 (2013).

[6] W. Zhang, T. Yoshida, X.J. Tang, and Q. Wang, Text clustering using frequent itemsets. Knowledge-Based Systems, **23**, 379-388 (2010).

[7] F. Saeed, N. Salim, and A. Abdo, Information theory and voting based consensus clustering for combining multiple clusterings of chemical structures. Molecular Informatics, **32**, 591-598 (2013).

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

[9] A.K. Jain, and R.C. Dubes, Algorithms for clustering data, Prentice Hall, New Jersey, 1988.

[10] J. Han, M. Kamber, and J. Pei, Data mining concepts and techniques. 3rd, Morgan Kaufmann, Massachusetts, 2012.

[11] Z. Huang, Clustering large data sets with mixed numeric and categorical values. In the first Pacific-Asia Conference on Knowledge Discovery and Data Mining. (1997).

[12] A. Ahmad, and L. Dey, A k-mean clustering algorithm for mixed numeric and categorical data. Data & Knowledge Engineering, **63**, 503-527 (2007).

[13] J. Ji, T. Bai, C. Zhou, C. Ma, and Z. Wang, An improved k-prototypes clustering algorithm for mixed numeric and categorical data. Neurocomputing, **120**, 590-596 (2013).

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

[15] L. Bai, J. Liang, and C. Dang, An initialization method to simultaneously find initial cluster centers and the number of clusters for clustering categorical data. Knowledge-Based Systems, **24**, 785-795 (2011).

[16] F. Cao, J. Liang, and L. Bai, A new initialization method for categorical data clustering. Expert Systems with Applications, **36**, 10223-10228 (2009).
[17] S. Wu, Q. Jiang, and J.Z. Huang. A new initialization method for clustering categorical data. In the 11th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), (2007).
[18] J. MacQueen, Some methods for classification and analysis of multivariate observations. In the fifth Berkeley Symposium on Mathematical Statistics and Probability, (1967).
[19] E.W. Forgy, Cluster analysis of multivariate data: efficiency vs interpretability of classifications. Biometrics, **21**, 768-769 (1965).
[20] R.C. Jancey, Multidimensional group analysis. Australian Journal of Botany, **14**, 127-130 (1966).
[21] P.S. Bradley, and U.M. Fayyad. Refining initial points for k-means clustering. In the 15th International Conference on Machine Learning. Morgan Kaufmann Publishers Inc, San Francisco, 1998.
[22] J.T. Tou, and R.C. Gonzalez, Pattern recognition principles, Addison-Wesley, Boston, 1974.
[23] T.F. Gonzales, Clustering to minimize the maximum intercluster distance. Theoretical Computer Science, **38**, 293-306 (1985).
[24] I. Katsavounidis, C.-C.J. Kuo, and Z. Zhang, A new initialization technique for generalized lloyd iteration. IEEE Signal Processing Letters, **1**, 144-146 (1994).
[25] M.B. A1-Daoud, and S.A. Roberts, New methods for the initialisation of clusters. Pattern Recognition Letters, **17**, 451-455 (1996).
[26] M.B. Al-Daoud, A new algorithm for cluster initialization. World Academy of Science, Engineering and Technology, **4**, 74-76 (2005).
[27] D. Arthur, and S. Vassilvitskii, k-means++ : the advantages of careful seeding. In Proceedings of the 18th Annual ACM-SIAM symposium on Discrete algorithms, 1027-1035 (2007).
[28] S.S. Khan, and A. Ahmad, Cluster center initialization algorithm for k-means clustering. Pattern Recognition Letters, **25**, 1293-1302 (2004).
[29] S.J. Redmond, and C. Heneghan, A method for initialising the k-means clustering algorithm using kd-trees. Pattern Recognition Letters, **28**, 965-973 (2007).
[30] F. Cao, J. Liang, and G. Jiang, An initialization method for the k-means algorithm using neighborhood model. Computers and Mathematics with Applications, **58**, 474-483 (2009).
[31] B. Yi, H. Qiao, F. Yang, and C. Xu, An improved initialization center algorithm for k-means clustering. In 2010 International Conference on Computational Intelligence and Software Engineering (CiSE),1-4 (2010).
[32] V. Kumar, J. Chhabra, and D. Kumar, Initializing cluster center for k-means using biogeography based optimization. In Advances in Computing, Communication and Control. Springer, Berlin, 448-456 (2011).
[33] S.S. Khan, and A. Ahmad, Cluster center initialization algorithm for k-modes clustering. Expert Systems with Applications, **40**, 7444-7456 (2013).
[34] L. Bai, J. Liang, C. Dang, and F. Cao, A cluster centers initialization method for clustering categorical data. Expert Systems with Applications, **39**, 8022-8029 (2012).
[35] Y. Sun, Q.M. Zhu, and Z.X. Chen, An iterative initial-points refinement algorithm for categorical data clustering. Pattern Recognition Letters, **23**, 875-884 (2002).
[36] S.S. Khan, and A. Ahmad. Computing initial points using density based multiscale data condensation for clustering categorical data. In the 2nd International Conference on Applied Articial Intelligence. (2003).
[37] P. Mitra, C.A. Murthy, and S.K. Pal, Density-based multiscale data condensation. IEEE Transactions on Pattern Analysis and Machine Intelligence, **24**, 734-747 (2002).
[38] Z. He, Farthest-point heuristic based initialization methods for k-modes clustering. CoRR, 2006. abs/cs/0610043.
[39] S.S. Khan, and S. Kant. Computation of initial modes for k-modes clustering algorithm using evidence accumulation. In the 20th international joint conference on Artifical intelligence, Morgan Kaufmann Publishers Inc, San Francisco, 2007.
[40] Z.X. Huang, and M.K. Ng, A fuzzy k-modes algorithm for clustering categorical data. IEEE Transactions on Fuzzy Systems, **7**, 446-452 (1999).
[41] W.M. Rand, Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association, **66**, 846-850 (1971).

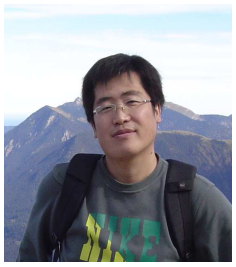
Jinchao Ji received the PhD degree from the College of Computer Science and Technology, Jilin University, in 2013. He is the author or co-author of more than ten scientific papers. His current research interests include diffuse of influence, social network analysis, artificial intelligence, and machine learning.

Wei Pang received the PhD degree in computing science from the University of Aberdeen in 2009. He currently holds a research fellow post in the University of Aberdeen. His research interests include qualitative reasoning, evolutionary algorithms, and artificial immune systems.

Yanlin Zheng received the PhD degree in Intelligent Information System from Tokushima University, Japan. She is a professor of Educational Technology in Northeast Normal University. Her research interests include big data supported educational evaluation and learning analytics.

Zhiqiang Ma received the Ph.D. degree from the College of Computer Science and Technology, Jilin University, China, in 2009. He is currently a professor of Computer Science and Information Technology in Northeast Normal University. His research interests include bioinformatics, data mining, software engineering.

Zhe Wang received the Ph.D. degree from the College of Computer Science and Technology, Jilin University, China, in 2005. He is currently an associate professor in Jilin University. His research interests include pattern recognition, social network analysis, artificial intelligence, clustering analysis and machine learning.



Libiao Zhang received the PhD degree in computer application and technology from the College of Computer Science and Technology of Jilin University in 2007. Currently, he is an associate professor of computer Science and Information Technology in Northeast Normal University. His research interests include computational intelligence, pattern recognition, machine learning, and internet of things.