International Journal of Innovative Computing, Information and Control
Volume **10**, Number **5**, October **2014**

ICIC International ©2014 ISSN 1349-4198
pp. 1873–1883

AN INITIALIZATION METHOD OF *K*-MEANS CLUSTERING ALGORITHM FOR MIXED DATA

Taoying Li, Zhihong Jin, Yan Chen and Angelo Dan Menga Ebonzo

Transportation Management College
Dalian Maritime University
No. 1, Linghai Road, Dalian 116026, P. R. China
ytaoli@126.com

Received November 2013; revised March 2014

Abstract. *The k-means clustering algorithm is undoubtedly the most widely used partitional algorithms. Unfortunately, due to its gradient descent nature, this algorithm is highly sensitive to the initialization of clustering. Initialization methods have been proposed to address this problem. In this paper, we present an overview of initialization methods of clustering for numerical data and categorical data respectively with an emphasis on their computational efficiency. We then propose a new initialization method for mixed data, which can obtain the good initial cluster centers using the MaxAvg distance, and give the effective k-means clustering for mixed data. Finally, the proposed method is verified on three different real world datasets from UCI Machine Learning Repository, and it is shown that the proposed method is effective and efficient for initializing and partitioning mixed data.*
Keywords: Mixed data, MaxAvg distance, Initialization of clustering, k-means clustering

1. **Introduction.** Clustering is one of the most important tasks in exploratory data analysis [1,2]. Its primary goals are to group the similar patterns into the same cluster and discover the meaningful structure of the data. Clustering has a long and rich history in a variety of scientific disciplines including anthropology, biology, medicine, psychology, statistics, mathematics, engineering, and computer science [3].

Clustering algorithms can be broadly classified into five groups: hierarchical, partitioning, density-based, grid-based and model-based clustering. Hierarchical and partitioning clustering algorithms are used widely in practice [4]. Hierarchical algorithm can be further divided into bottom-up and top-down algorithms. Traditional partition clustering algorithms include k-means, and k-modes. The k-means clustering is undoubtedly the most widely used partitional clustering algorithm, and its sensitivity of initialization of clustering has captured the attention of the clustering communities for quite a long time.

The rest of the paper is organized as follows. Section 2 presents some related works to the initialization methods for numerical data and categorical data. Section 3 describes the initialization of clustering for mixed data, which is our proposed methodology. Section 4 presents the experimental results for UCI dataset. Conclusion and remark are given in Section 5.

2. **Related Work.** In this section, we review some common studies on initialization of clustering.

2.1. **Review of initializing clustering for numerical data.** Celebi et al. [5] compared eight commonly used linear time complexity initialization methods on a large and diverse collection of data sets using various performance criteria and found that popular initialization methods often perform poorly and that are strong alternatives to methods. Nie et al. [6] proposed an actively self-training clustering method, in which the samples were actively selected as training set to minimize an estimated Bayes error, and then explored semi-supervised learning to perform clustering. Reddy and Jana [7] presented a novel method that selected the initial cluster centers with the help of Voronoi diagram constructed from the given set of data points. The initial cluster centers were effectively selected from those points which lied on the boundary of higher radius Voronoi circles. Naik et al. [8] advanced the TLBO (Teaching Learning Based Optimization) to address the problem of initializing centers of clusters, in which the search space of given dataset was used to find out near-optimal cluster centers and taken use of reformulated c-mean objective function to evaluate centers. Redmond and Heneghan's method [9] first constructs a kd-tree of the data points to perform density estimation and then uses a modified maximum method to select K centers from densely populated leaf buckets. The computational cost of this method is dominated by the complexity of kd-tree construction, which is $O(N \log N)$. Cao et al. [10] formalized Astrahan's density based method within the framework of a neighborhood-based rough set model. In this model, the e-neighborhood of a point is defined as the set of points within e distance from it according to a particular distance measure. Based on this neighborhood model, the concepts of cohesion and coupling are defined. The former is a measure of the centrality of a point with respect to its neighborhood, whereas the latter is a measure of separation between two neighborhoods. The method first sorts the data points in decreasing order of their cohesion and takes the point with the greatest cohesion as the first center. It then traverses the points in sorted order and takes the first point that has a coupling of less than e with the previously selected centers as the ith ($i = \{2, 3, \ldots, k\}$) center. The computational cost of this method is dominated by the complexity of the e-neighborhood calculations, which is $O(N^2)$. Li and Wang [11] assume that there is at least one dense subset of data in a cluster; and the dense subsets between different clusters are more distant than those in the same cluster. A minimum spanning tree is built for the given data set. The dense subsets can be found through the search from root trees, and their densities are obtained by the estimation technique for data density. The initial cluster centers are picked out from the dense subsets that are dense enough and distant enough from each other. Zhong and Zhang [12] compared the influence of five initialization methods on unsupervised classification algorithms respectively by means of various experiments in remote sensing images.

2.2. **Review of initializing clustering for categorical data.** Currently, most methods of initialization cluster centers are mainly for numerical data. However, these methods used in cluster centers initialization for numerical data are not applicable to categorical data due to a lack of geometry for the categorical data. In this paper, we propose a new methodology to overcome this issue. Bai et al. [13] proposed a novel initialization method for categorical data, in which the distance and the density were adapted together to select initial centers and overcame shortcomings of the existing initialization methods for categorical data. In [14], they could not only obtain the good initial cluster centers but also provide a criterion to find candidates for the number of clusters ([14]). Sun et al. [15], introduced 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 ([16]) to the k-modes clustering, but (its time cost is high and) the parameters of this method

are plenty and need to be asserted in advance [16]. Arthur and Vassilvitskii [17] proposed a careful seeding for initial cluster centers to improve clustering results. However, due to a lack of intuitive geometry for categorical data, the techniques used in cluster centers initialization for numerical data are not applicable to categorical data. Huang [18] suggested to select the first k distinct objects from the data set as the initial k modes or to assign the most frequent categories equally to the initial k modes. Though the methods are to make the initial modes diverse, a uniform criterion is not given for selecting k initial modes. Barbara et al. [19] used the MaxMin distances method to find the k most dissimilar data objects from the data set as initial seeds in Coolcat algorithm. However, the method only considers the distance between the data objects, by which outliers may be selected. The distance and the density were integrated together to propose a cluster centers initialization method respectively in [20,21]. The difference between the two methods is the definition of the density of an object and Cao et al. defined the density of an object based on frequency of attribute values [20]. Besides, there are many other researches on initializing partition, such as specifying initial values by finding a large number of local modes and then obtaining representatives from the most separated ones [22].

3. Proposed Methodology.

In this section, we first present the general method of DCD (Distance for Categorical Data), and then introduce our proposed DCD.

3.1. Distance methods for categorical data.

Huang [23] presented a cost function for measuring the distance of clustering for mixed data, and it can be shown in (1).

$$\vartheta(d_j, c_l) = \sum_{i=1}^{m_r} (x_{ji} - c_{li})^2 + \mu \sum_{i'=m_r+1}^{m} \delta(x_{ji'}, c_{li'})^2 \tag{1}$$

Here, m_r stands for the number of numerical attributes. Suppose that numerical attributes i (could) start from 1 to m_r and categorical attributes i' (could) start from $m_r + 1$. The parameter μ represents the correlation coefficient, and its range is $(0, 1]$. In previous researches, it is supposed that: $\delta(p, q) = 0$ when $p = q$ and $\delta(p, q) = 1$ when $p \neq q$.

In this research, we prove that the distance cannot be expressed clearly by $\delta(p, q) = 1$ when $p \neq q$. We then, introduce the relationship degree n into the distance.

The relationship degree (RD) is definite in this paper as follows: suppose that x_j and x_t are two arbitrary data objects, including m_c categorical attributes, and $RD(x_j, x_t)$ is number that $x_{ji} = x_{ti}$ when $i = 1, 2, \ldots, m$.

From the definition of relationship degree, we know that the maximum value of relationship degree is the number of categorical attributes. Therefore, we state Theorem 3.1 as below:

Theorem 3.1. *Suppose that x_j and x_t are two arbitrary data objects, including m_c categorical attributes, if $RD(x_j, x_t) = 0$, then for any object x_s, satisfying $RD(x_j, x_s) = a$ and $RD(x_t, x_s) = b$ $(a, b > 0)$, $RD(x_j, x_t)$ can be modified as $RD(x_j, x_t) = \min(a, b)/2$.*

The proof of Theorem 3.1 will be given after that for the definition for DCD.

Distance for Categorical Data (DCD): Let x_j and x_t be any two objects, including m_c categorical attributes, and then the distance $\delta(x_j, x_t)$ between x_j and x_t is given as in (2).

$$\delta(x_j, x_t) = m_c - RD(x_j, x_t) \tag{2}$$

Proof: In what follows, we prove that δ is a distance, since it satisfies below of distance space characteristics.

1) Reflexivity:

$$\delta \geq 0, \text{ and } \delta(x_j, x_t) = 0 \Leftrightarrow x_j = x_t \tag{3}$$

2) Symmetry:

$$\delta(x_j, x_t) = \delta(x_t, x_j) \tag{4}$$

3) Transitivity:

$$\delta(x_j, x_t) \leq \delta(x_t, x_s) \text{ and } \delta(x_t, x_s) \leq \delta(x_t, x_s) \Rightarrow \delta(x_j, x_t) \leq \delta(x_t, x_s) \tag{5}$$

4) Triangle theorem:

$$\delta(x_j, x_t) \leq \delta(x_j, x_s) + \delta(x_s, x_t) \tag{6}$$

It is obvious that the reflexivity, symmetry and transitivity are established, and we only need to prove the triangle theorem(s).

Proof for triangle theorems: Let x_j, x_s and x_t be any three different objects, and supposing all of them have m_c categorical attributes, if $RD(x_j, x_t) = 0$, $RD(x_j, x_s) = a$ and $RD(x_t, x_s) = b$, then $RD(x_j, x_t) = \min(a, b)/2$, and this distance can be updated as follows.

$\delta(x_j, x_t) = m_c - RD(x_j, x_t) = m_c - \min(a, b)/2 < m_c$

$\delta(x_t, x_s) + \delta(x_t, x_s) = (m_c - RD(x_t, x_s)) + (m_c - RD(x_t, x_s)) = (m_c - a) + (m_c - b) = m_c + (m_c - a - b)$

Due to $RD(x_j, x_t) = 0$, $RD(x_j, x_s) = a$ and $RD(x_t, x_s) = b$, we can get $(m_c - a - b) \geq 0$, so $\delta(x_t, x_s) + \delta(x_t, x_s) \geq m_c$.

According to the proof mentioned above, we know that it is a distance space and could be used to obtain the distance between any two categorical data.

Proof for Theorem 3.1: Let $RD(x_j, x_t) = x$, we know that $(m_c - a) + (m_c - b) > (m_c - x)$ due to triangle theorems, which can be shown in Figure 1. Next we can grasp $x > a + b - m_c$, at the same time to meet $|(m_c - a) - (m_c - b)| > (m_c - x)$. Supposing $b > a$, we obtain $x < m_c - (b - a)$.

From the definition of relationship degree, we can establish the structure of relationship degree based on $RD(x_j, x_t) = 0$, $RD(x_j, x_s) = a$ and $RD(x_t, x_s) = b$, which can be shown as Figure 2.

Then we know that relationship degree satisfies the characteristics of geometric structure from Figure 2.

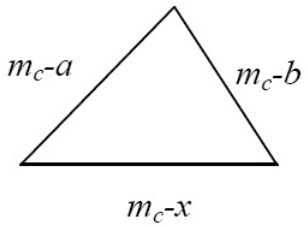

FIGURE 1. Triangle theorems

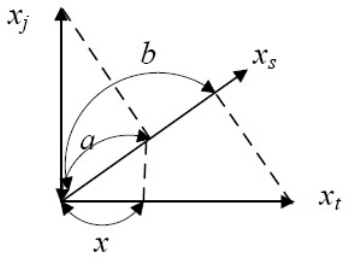

FIGURE 2. Structure of relationship degree among objects

$b/m_c = x/a$, can be transformed to $m_c = ab/x$, substituting $m_c = ab/x$ into $x > a+b-m_c$, we get $x > a+b-ab/x$, and we know that $x <$ a according to $x < m_c-(b-a)$.

Thus, in this paper, we let $x = a/2$. So $RD(x_j, x_t) = \min(a, b)/2$ meets the requirements.

We will give an example for relationship degree in Table 1.

TABLE 1. Categorical data

Object	$a1$	$a2$	$a3$	$a4$
x_j	A1	B1	C1	D1
x_s	A1	B1	C2	D2
x_t	A2	B2	C2	D2

From Table 1, we know that $RD(x_j, x_s) = 2$ and $RD(x_t, x_s) = 2$, and then we can get the distance $RD(x_j, x_t) = 0$ by traditional function and $RD(x_j, x_t) = 2/2 = 1$ by relationship degree mentioned above.

3.2. **Distance methods for mixed data.** In this paper, we define the distance for mixed data (DMD) as follows:

$$\vartheta(x_t, x_j) = \sum_{i=1}^{m-m_c} (x_{ti} - x_{ji})^2 + \mu\delta(x_t, x_j) \tag{7}$$

Here, x_j, x_t are two objects including m_c categorical attributes $(m - m_c)$ numerical attributes, and $\vartheta(x_j, x_t)$ is the distance between x_j and x_t. From the equation, we know that the DMD includes two parts, the former one is the Euclidean distance of x_j and x_t for numerical data, and the latter one is their DCD for categorical data, and μ reflects the impact of categorical attributes.

Dataset should be normalized before they were used in order to be dimensionless. We adopt the Min-Max method, which can be shown as (8).

$$x'_{ji} = \frac{x_{ji} - \min\limits_t x_{ti}}{\max\limits_t x_{ti} - \min\limits_t x_{ti}} \tag{8}$$

Then, we obtain that the value of x'_{ji} will be in $[0, 1]$ for any j.

We also add some parameters into (7) while the influence of attributes for results of clustering is different. Then we use the new weighted distance for mixed data (WDMD) as (8).

$$\vartheta(x_t, x_j) = \sum_{i=1}^{m-m_c} \omega_i(x_{ti} - x_{ji})^2 + \mu\delta(x_t, x_j) \tag{9}$$

Here,

$$\delta(x_t, x_j) = \sum_{i=m-m_c+1}^{m} \omega_i - \sum_{i=m-m_c+1}^{m} \omega_i RD(x_{ti}, x_{ji}) \tag{10}$$

where ω_i denotes the impact of the ith attribute for clustering, and $\sum\limits_{i=1}^{m} \omega_i = 1$.

3.3. **Initialization of clustering for mixed data.** Above all, we integrate methods of density and grid to get high-density sub area, and then use the method of MaxAvg distance to address the issue of initializing the centers of the clusters, in which we explore the DMD distance.

MaxAvg distance is presented in (11).

$$c_l = \left\{ x_t | Avg(l-1,t) = \max_{j=1\to h} Avg(l-1,j) \right\} \tag{11}$$

where

$$Avg(l-1,j) = \frac{1}{l-1} \sum_{i=1}^{l-1} \vartheta(c_i, x_j) \tag{12}$$

Here, $l = 2, 3, \ldots, k$ and $j = 1, 2, 3, \ldots, h$. l is the number of initial centers that we have obtained and x_j is the representative point of the jth grid. c_l is the lth center and meets $c_l \neq c_{l+1}$ for any l. The sum of distance among x_j and all centers can be shown in Figure 4.

The computation of the initialization can be shown as follows:

Input: m, μ, k, x, and m is the number of attributes, μ is impact factor, k is the number of cluster desired, x is dataset.

Output: $C = \{c_1, c_2, \ldots, c_k\}$.

Let $l = 1$ and m_i denotes the number of values of the ith attribute for categorical data.

Begin

  For $l = 1$ to k

    For $i = 1$ to m

      If $l == 1$ && the ith attribute is categorical

        c_{li} = value that turns up most frequently in the ith attribute.

      Else If $l == 1$ && the ith attribute is numerical

        $c_{li} = \Sigma_j x_{ji}/n$

      Else

        Get the c_l according to (11) and (12).

      End If

    End For

  End For

End

3.4. **The *k*-means clustering for mixed data.** In this section, we adopt the objective function of clustering as shown in (13).

$$\min F(\omega, \tau, c) = \sum_{j=1}^{n} \tau_{lj} \left(\sum_{i=1}^{m-mc} w_{li}(c_{li} - x_{ji})^2 + \mu \vartheta(c_l, x_j) \right) \tag{13}$$

where $\sum_{l=1}^{k} \tau_{lj} = 1$, $1 \leq j \leq n$, $\tau_{lj} \in \{0, 1\}$, $\sum_{i=1}^{m} \omega_{li} = 1$, $0 \leq \omega_{li} \leq 1$, $1 \leq l \leq k$.

Here, if $\tau_{lj} = 1$, it means that the jth object belongs to the lth cluster, and vice versa.

Next, we modify the centers in order to get better results.

Input: T, s and T is the times of iteration.
Output: $C = \{c_1, c_2, \ldots, c_k\}$.
Let A be the set that includes all values of some categorical attributes and a_h denotes the number of the hth value of the ith attribute.

```
Begin
    For t = 1 to T
      For j = 1 to n
        If ϑ(c_l, x_j) ≤ ϑ(c_z, x_j) for any z (1 ≤ z ≤ k)
          τ_lj = 1
        Else
          τ_lj = 0
        End If
      End For
      For l = 1 to k
        For i = 1 to m
          If the ith attribute is numerical
              c_li = 1
              y = 0
            For j = 1 to n
              c_li = c_li + τ_lj x_li
            End For
              c_li = c_li/Στ_lj
          Else
              A = ∅
              a_h = 0 for each h
              h = 0
            For any τ_lj = 1
              If x_ji ∉ A
                h = h + 1
                A_h = {x_ji}
                a_h = a_h + 1
              Else If x_ji ∈ A
                ∃A_r = x_ji
                a_r = a_r + 1
              End If
            End For
                c_li = A_r if a_r is max
          End If
        End For
      End For
    End For
End
```

4. **Experimental Studies.** In this study, we use the Diagnosis dataset, Iris dataset from the common UCI datasets to check our proposed methodology.

4.1. **Diagnosis dataset.** In this experiment, we suppose that all data are categorical and study two parts of records. We choose lumbar pain {No, Yes} and urine pushing

{No, Yes} as the decision attributes for the first sample data and occurrence of nausea {Yes, Yes} and lumbar pain {No, Yes} as the decision attributes for the second sample data. From their values we use the initialization of clustering for mixed data and the k-means clustering for mixed data proposed in this paper, the results are shown in Table 2.

TABLE 2. Inflammation of urinary bladder dataset

No	Sample 1	Sample 2
First attribute	lumbar pain	occurrence of nausea
Second attribute	urine pushing	lumbar pain
Range of value	{No, Yes}	{No, Yes}
Number of data	60	60
Initial centers	⟨Yes, No⟩ ⟨No, Yes⟩	⟨Yes, Yes⟩ ⟨No, Yes⟩
Final centers	⟨Yes, No⟩ ⟨No, Yes⟩	⟨Yes, Yes⟩ ⟨No, No⟩
Times of clustering of proposed method	1	2
Times of clustering of traditional method	5	8

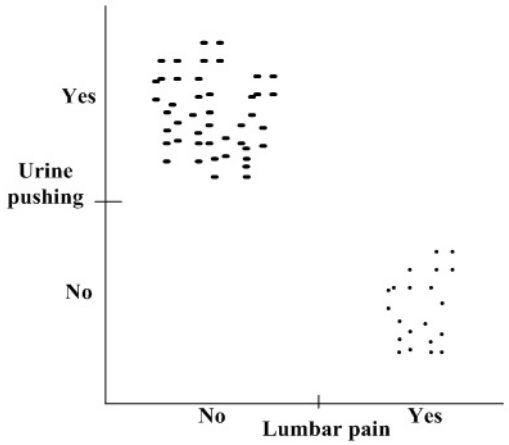

FIGURE 3. Results for Sample 1

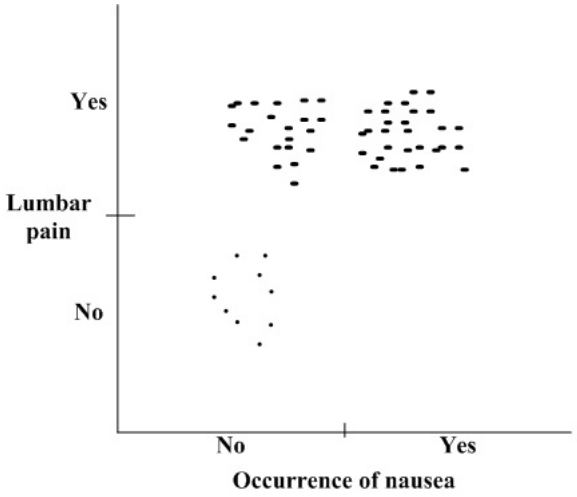

FIGURE 4. Results for Sample 2

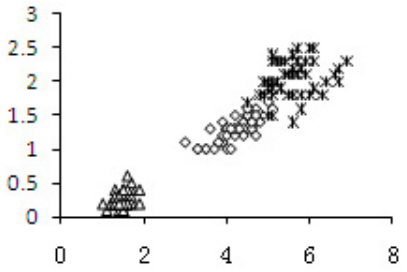

FIGURE 5. Results for Iris dataset

Therefore, the method of initialization proposed in this paper is effective.

Based on the analyses presented in the previous section, we know that the traditional methods of initialization for single numerical or categorical data are ineffective and unreliable since they often result in slower k-means convergence. The methods proposed in this paper produce very good initial clustering (see Table 2, Figure 3 and Figure 4), which makes it possible to use them as initial algorithms for categorical data.

4.2. **Iris dataset.** Here, we will verify the proposed methods for numerical data using Iris dataset, which includes 4 numerical attributes and totally 150 objects. Figure 5 gives the results of clustering using proposed methods in three clusters and its precision can be shown in Table 3.

TABLE 3. Precision for Iris dataset

Algorithms	Random initialization	Proposed initialization
Traditional k-means	88%	91.67%
EWKM algorithm [4]	90.33%	94%
IWEKM algorithm [24]	91.67%	95.33%
Proposed method	91.33%	95.67%

From Table 3, we know that the results of proposed methods have higher efficiency and precision than that of traditional k-means clustering.

4.3. **Zoo dataset.** Here, we verify the proposed methodology using Zoo dataset with 16 attributes, and its results are shown in Table 4, from which we can see that the proposed methodology is more effective.

TABLE 4. Precision for Zoo dataset

Algorithms	Random initialization	Proposed initialization
Squeezer	79.2%	81.2%
GAClust	77.2%	79.2%
ccdByEnsemble	75.2%	76.2%
CEMD	88.1%	91.1%
EWKM	85.1%	87.1%
IWEKM	84.1%	88.1%
Proposed method	89.1%	91.1%

5. **Conclusions.** Mixed data are ubiquitous in real-world databases, especially in decision making databases. However, its performance of clustering algorithms strongly depends on an initial set of cluster centers. In this paper we presented an overview of initialization methods of clustering for numerical data and categorical data respectively with an emphasis on their computational efficiency. We then proposed a new initialization method for mixed data and gave its process of computation. Next, we presented the k-means clustering for mixed data, including its computation. Finally, we tested the proposed method using diagnosis dataset, Iris dataset and Zoo dataset, real world datasets from UCI Machine Learning Repository. The experimental results were analyzed and illustrated that the proposed method is effective and efficient for initializing and classifying mixed data.

Acknowledgment. This work is partially supported by the National Natural Science Foundation of China (Grant No. 71271034), General Project of Liaoning Provincial Education Department (Grant No. L2012173), Liaoning Academy of social science fund project (Grant No. L13DGL060) and the Fundamental Research Funds for the Central Universities (Grant No. 3132014080). The authors are very grateful to the editors and reviewers for their valuable comments and suggestions.

REFERENCES

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

[2] S. R. Kannan, S. Ramathilagam and P. C. Chung, Effective fuzzy c-means clustering algorithms for data clustering problems, *Expert Systems with Applications*, vol.39, no.7, pp.6292-6300, 2012.

[3] A. K. Jain, Data clustering: 50 years beyond k-means, *Pattern Recognition Letters*, vol.31, no.8, pp.651-666, 2010.

[4] L. Jing, M. K. Ng and J. Z. Huang, An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data, *IEEE Transactions on Knowledge and Data Engineering*, vol.19, no.8, pp.1026-1041, 2007.

[5] 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*, vol.40, no.1, pp.200-210, 2013.

[6] F. Nie, D. Xu and X. Li, Initialization independent clustering with actively self-training method, *IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics*, vol.42, no.1, pp.17-27, 2012.

[7] D. Reddy and P. K. Jana, Initialization for k-means clustering using Voronoi diagram, *Procedia Technology*, vol.4, pp.395-400, 2012.

[8] A. Naik, S. C. Satapathy and K. Parvathi, Improvement of initial cluster center of c-means using teaching learning based optimization, *Procedia Technology*, vol.6, pp.428-435, 2012.

[9] S. J. Redmond and C. Heneghan, A method for initializing the k-means clustering algorithm using kd-trees, *Pattern Recognition Letters*, vol.28, no.8, pp.965-973, 2007.

[10] F. Cao, J. Liang and G. Jiang, An initialization method for the k-means algorithm using neighborhood model, *Computers and Mathematics with Applications*, vol.58, no.3, pp.474-483, 2009.

[11] C. Li and Y. Wang, New initialization method for cluster center, *Control Theory and Applications*, vol.27, no.10, pp.1435-1440, 2010.

[12] Y. Zhong and L. Zhang, Initialization methods for remote sensing image clustering using k-means algorithm, *Systems Engineering and Electronics*, vol.32, no.9, pp.2009-2014, 2010.

[13] L. Bai, J. Liang, C. Dang and F. Cao, A cluster centers initialization method for clustering categorical data, *Expert Systems with Applications*, vol.39, no.9, pp.8022-8029, 2012.

[14] 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*, vol.24, no.6, pp.785-795, 2011.

[15] Y. Sun, Q. M. Zhu and Z. X. Chen, An iterative initial-points refinement algorithm for categorical data clustering, *Pattern Recognition Letters*, vol.23, no.7, pp.875-884, 2002.

[16] S. Bradley and U. M. Fayyad, Refining initial points for k-means clustering, *Proc. of the 15th International Conf. on Machine Learning*, Morgan Kaufmann, San Francisco, CA, USA, pp.91-99, 1998.

[17] D. Arthur and S. Vassilvitskii, K-means++: The advantages of careful seeding, *Proc. of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, New Orleans, LA, USA, pp.1027-1035, 2007.

[18] Z. X. Huang, Extensions to the k-means algorithm for clustering large data sets with categorical values, *Data Mining Knowledge Discovery*, vol.2, no.3, pp.283-304, 1998.

[19] D. Barbara, J. Couto and Y. Li, COOLCAT: An entropy-based algorithm for categorical clustering, *Proc. of the 11th International Conference on Information and Knowledge Management*, McLean, VA, pp.582-589, 2002.

[20] F. Y. Cao, J. Y. Liang and L. Bai, A new initialization method for categorical data clustering, *Expert Systems with Applications*, vol.33, no.7, pp.10223-10228, 2009.

[21] S. Wu, Q. S. Jiang and Z. X. Huang, A new initialization method for categorical data clustering, *Lecture Notes in Computer Science*, vol.4426, pp.972-980, 2007.

[22] R. Maitra, Initializing partition-optimization algorithms, *IEEE ACM Transactions on Computational Biology and Bioinformatics*, vol.6, no.1, pp.144-157, 2009.

[23] Z. Huang, Clustering large datasets with mixed numeric and categorical values, *Proc. of the 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining,* World Scientific, Singapore, pp.21-34, 1997.

[24] T. Li and Y. Chen, A weight entropy k-means algorithm for lustering dataset with mixed numeric and categorical data, *Proc. of the 5th International Conference on Fuzzy Systems and Knowledge Discovery*, Jinan, China, pp.36-41, 2008.

[25] Y. T. Kao, E. Zahara and I. W. Kao, A hybirdized approach to data clustering, *Expert Systems with Applications*, vol.34, no.3, pp.1754-1762, 2008.