Expert Systems with Applications 36 (2009) 10223–10228

Contents lists available at ScienceDirect

Expert Systems with Applications

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

A new initialization method for categorical data clustering

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

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

ARTICLE INFO

Keywords:
Density
Distance
Initialization method
Initial cluster center
k-modes algorithm

ABSTRACT

In clustering algorithms, choosing a subset of representative examples is very important in data set. Such "exemplars" can be found by randomly choosing an initial subset of data objects and then iteratively refining it, but this works well only if that initial choice is close to a good solution. In this paper, based on the frequency of attribute values, the average density of an object is defined. Furthermore, a novel initialization method for categorical data is proposed, in which the distance between objects and the density of the object is considered. We also apply the proposed initialization method to *k*-modes algorithm and fuzzy *k*-modes algorithm. Experimental results illustrate that the proposed initialization method is superior to random initialization method and can be applied to large data sets for its linear time complexity with respect to the number of data objects.

1. Introduction

Clustering data based on a measure of similarity is a critical step in scientific data analysis and in engineering systems. A common method is to use data to learn a set of centers such that the sum of squared errors between objects and their nearest centers is small (Brendan & Delbert, 2007). At present, the popular partition clustering technique usually begins with an initial set of randomly selected exemplars and iteratively refines this set so as to decrease the sum of squared errors. Due to the simpleness, random initialization method has been widely used. However, these clustering algorithms need to be rerun many times with different initializations in an attempt to find a good solution. Furthermore, random initialization method works well only when the number of clusters is small and chances are good that at least one random initialization is close to a good solution. Therefore, how to choose initial cluster centers is extremely important as they have a direct impact on the formation of final clusters. Based on the difference in data type, selection of initial cluster centers mainly can be classified into numeric data and categorical data. Aiming at numeric data, several attempts have been reported to solve the cluster initialization problem (Bradley, Mangasarian, & Street, 1997; Bradley & Fayyad, 1998; Duda & Hart, 1973; Fisher, 1987, 1996; Higgs, Bemis, Watson, & Wikel, 1997; Milligan, 1980; Meila & Heckerman, 1998; Pen, Lozano, & Larraaga, 1999, 2004; Snarey, Terrett, Willet, & Wilton, 1997; Ward, 1963). However, to date, few researches are concerned with initialization of categorical data. Huang (1998) introduced two initial mode selection methods for *k*-modes algorithm. The first method selects the first *k* distinct objects from the data set as the initial *k*-modes. The second method assigns the most frequent categories equally to the initial *k*-modes. Though the second method is to make the initial modes diverse, a uniform criterion is not given for selecting initial *k*-modes yet. Sun, Zhu, and Chen (2002) introduced an initialization method that 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. Wu, Jiang, and Huang (2007) proposed the new initialization method that 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 cannot be guaranteed. In summary, there are no universally accepted methods for selecting initial cluster centers currently. Hence, it is very necessary to propose a new initialization method for categorical data.

For numeric data, Higgs et al. (1997) and Snarey et al. (1997) suggested using a MaxMin algorithm in order to select a subset of the original database as the initial centroid to establish the initial clusters. However, in MaxMin algorithm, the first cluster center is still randomly selected, and selection of the rest of the cluster centers only limits the distance between the object and existing cluster centers. In this paper, the average density of an object is defined, based on the frequency of attribute values. The object with the maximum density is taken as the first cluster center. For the rest of the cluster centers, we extend the MaxMin algorithm and take into account not only the distance between the objects, but

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

0957-4174/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2009.01.060

also the density of the object. Based on the above ideas, a novel initialization method for categorical data is proposed. The proposed initialization method is used along with k-modes algorithm and fuzzy k-modes algorithm with different dissimilarity measures, respectively. The time complexity of initialization method is analyzed. Comparisons with random initialization method illustrate the effectiveness of this approach.

The outline of the rest of this paper is as follows. In Section 2, k-modes algorithm and fuzzy k-modes algorithm are introduced. In Section 3, a new initialization method for categorical data is proposed. In Section 4, the effectiveness of the new initialization method by comparative random initialization method is demonstrated. In Section 5, general discussion and the conclusions of this work are given.

2. The *k*-modes algorithm and the fuzzy *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 (Liang & Li, 2005; Pawlak, 1991; Zhang, Wu, Liang, & Li, 2001). Data in the real world usually contain categorical attributes. More formally, a categorical information system is described as quadruples $IS = (U, A, V, f)$, where

(1) U is the nonempty set of objects, which is 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 $x \in U$, $f(x, a) \in V_a$.

Let $IS = (U, A, V, f)$ be a categorical information system, then $x, y \in U$ and $a \in A$. The simple matching dissimilarity measure between x and y is defined as follows:

$$d(x,y) = \sum_{a\in A} \delta_a(x,y),$$

where

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

It is easy to verify that the function d defines a metric space on the set of categorical objects.

The k-modes algorithm (Huang, 1997a, 1997b) uses the k-means paradigm to cluster categorical data. The objective of clustering a set of $n = |U|$ objects into k clusters is to find W and Z that minimize

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

subject to

$$\omega_{li} \in \{0,1\}, \quad 1 \leqslant l \leqslant k, \quad 1 \leqslant i \leqslant n \tag{2}$$

$$\sum_{l=1}^{k} \omega_{li} = 1, \quad 1 \leqslant i \leqslant n \tag{3}$$

and

$$0 < \sum_{i=1}^{n} \omega_{li} < n, \quad 1 \leqslant l \leqslant k, \tag{4}$$

where $k(\leqslant n)$ is a known number of clusters, $W = [\omega_{li}]$ is a k-by-n {0,1} matrix, $Z = [z_1, z_2, \ldots, z_k]$, and z_l is the lth cluster center with the categorical attributes $a_1, a_2, \ldots, a_{|A|}$. The whole process of the k-modes algorithm is described as follows (Ng, Li, Huang, & He, 2007):

Step 1. Choose an initial point $Z^{(1)} \in \mathbb{R}^{|A|k}$. 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.

By modifying a simple matching dissimilarity measure for categorical objects, Ng et al. (2007) proposed a new dissimilarity measure, and given a rigorous proof that the object cluster membership assignment method and the mode updating formulae under the new dissimilarity measure indeed minimize the objective function. The new dissimilarity measure is defined as follows:

$$d_n(z_l, x_i) = \sum_{a\in A} \phi_a(z_l, x_i),$$

where

$$\phi_a(z_l, x_i) = \begin{cases} 1, & \text{if} \quad f(z_l, a) \neq f(x_i, a), \\ 1 - \frac{|c_{l,a}|}{|c_l|}, & \text{otherwise}. \end{cases}$$

where $|c_l|$ is the number of objects in lth cluster, and $|c_{l,a}|$ is the number of objects with category $f(z_l, a)$ of the ath attribute in the lth cluster.

Like the k-means algorithm, the k-modes algorithm also produces locally optimal solutions that are dependent on the initial modes and the order of objects in the data set. Huang (1998) introduced two initial mode selection methods. The first method selects the first k distinct records from the data set as the initial k-modes. Despite being used in a wide array of applications, clustering results are fluctuated with different initial cluster centers. The second method is implemented by calculating the frequencies of all categories for all attributes. The second method can lead to better clustering results; however, there has been no method so far to combine the category to form the appropriate mode.

In 1999, Huang and Ng (1999) proposed the fuzzy k-modes algorithm for clustering categorical objects based on the extensions to the fuzzy k-means algorithm. The objective of the fuzzy k-modes clustering is also to find W and Z that minimize

$$F_c(W,Z) = \sum_{l=1}^{k}\sum_{i=1}^{n} \omega_{li}^{\alpha} d(z_l, x_i) \tag{5}$$

subject to

$$\omega_{li} = \begin{cases} 1, & \text{if} \quad x_i = z_l, \\ 0, & \text{if} \quad x_i = z_l, \ h \neq l \\ 1 \Big/ \sum_{h=1}^{k} \left[\frac{d(z_l, x_i)}{d(z_h, x_i)}\right]^{1/(\alpha-1)} & \text{if} \quad x_i \neq z_l \quad \text{and} \quad x_i \neq z_h, \ 1 \leqslant h \leqslant k \end{cases} \tag{6}$$

and also subject to (3) and (4), where α is the weighting component, $W = (\omega_{li})$ is the $k \times n$ fuzzy membership matrix. Note that $\alpha = 1$ gives the hard k-modes clustering. This procedure removes the numeric-only limitation of the fuzzy k-means algorithm. Moreover, the fuzzy partition matrix provides more information to help the user to determine the final clustering and to identify the boundary objects. Such information is extremely useful in applications such as data mining in which the uncertain boundary objects are sometimes more interesting than objects which can be clustered with certainty. In 2009, Gan, Wu, and Yang (2009) presented a hybrid genetic fuzzy k-modes algorithm. However, the fuzzy k-modes algorithm is the same as the k-modes algorithm and is sensitive to initial cluster centers. To solve these problems, a new initialization method for categorical data is proposed in Section 3.

Table 1
The summary of clustering result of the *k*-modes on the soybean data.

The *k*-modes algorithm	Huang's dissimilarity measure		Ng's dissimilarity measure	
	Random	The proposed method	Random	The proposed method
AC	0.8564	1.0000	0.9364	1.0000
PR	0.9000	1.0000	0.9559	1.0000
RE	0.8402	1.0000	0.9257	1.0000
Iteration times	3.2800	4.0000	4.0700	2.0000

3. A new initialization method for categorical data

In this section, based on the average density of the object and the distance between objects, a new initialization method for categorical data is proposed and the time complexity of the proposed algorithm is analyzed.

To give the density of categorical data, firstly, we review mode of a set (Huang, 1998). A mode of $\mathbf{X} = \{x_1, x_2, \ldots, x_n\}$ is a vector $Q = [q_1, q_2, \ldots, q_m]$ that minimises $D(\mathbf{X}, Q) = \sum_{i=1}^{n} d(x_i, Q)$. In other words, $q_i (1 \leqslant i \leqslant m)$ is the most frequent value in $\mathbf{X}$ with respect to the ith attribute such that the vector Q is a mode. Here, Q is not necessarily an object of $\mathbf{X}$. When a mode is not an object of a set, it can be assumed a virtual object. Therefore, based on the above ideas, the average density of an object x in a given set is defined as follows.

Definition 1. Let $IS = (U, A, V, f)$ be a categorical information system. For any $x \in U$, the average density of x in U with respect to A is defined as

$$Dens(x) = \frac{\sum_{a \in A} Dens_a(x)}{|A|}$$

where $Dens_a(x)$ is the density of object x in U with respect to a, given by

Table 2
The summary of clustering result of the fuzzy *k*-modes on the soybean data.

The Fuzzy *k*-modes algorithm	Huang's dissimilarity measure		Ng's dissimilarity measure	
	Random	The proposed method	Random	The proposed method
AC	0.8443	1.0000	0.9204	1.0000
PR	0.8890	1.0000	0.9485	1.0000
RE	0.8277	1.0000	0.9074	1.0000
Iteration times	3.4900	3.0000	4.3400	3.0000

Table 3
The summary of clustering result of the *k*-modes on the zoo data.

The *k*-modes algorithm	Huang's dissimilarity measure		Ng's dissimilarity measure	
	Random	The proposed method	Random	The proposed method
AC	0.8356	0.8812	0.8577	0.8812
PR	0.8186	0.8702	0.8048	0.8492
RE	0.6123	0.6714	0.6556	0.6714
Iteration times	2.4500	3.0000	4.2000	4.0000

Table 4
The summary of clustering result of the fuzzy *k*-modes on the zoo data.

The Fuzzy *k*-modes algorithm	Huang's dissimilarity measure		Ng's dissimilarity measure	
	Random	The proposed method	Random	The proposed method
AC	0.8403	0.9208	0.8569	0.8812
PR	0.8241	0.8819	0.8421	0.8648
RE	0.6232	0.7857	0.6561	0.6714
Iteration times	2.6300	1.0000	3.4300	2.0000

Table 5
The summary of clustering result of the *k*-modes on the breast-cancer data.

The *k*-modes algorithm	Huang's dissimilarity measure		Ng's dissimilarity measure	
	Random	The proposed method	Random	The proposed method
AC	0.8461	0.9113	0.8503	0.8655
PR	0.8700	0.9292	0.8900	0.9069
RE	0.7833	0.8773	0.7864	0.8079
Iteration times	1.6800	3.0000	2.3500	3.0000

$$Dens_a(x) = \frac{|\{y \in U | f(x,a) = f(y,a)\}|}{|U|}$$

Obviously, we have $\frac{1}{|U|} \leqslant Dens(x) \leqslant 1$. For any $a \in A$, if $|\{y \in U | f(x,a) = f(y,a)\}| = 1$, then $Dens(x) = \frac{1}{|U|}$. If $|\{y \in U | f(x,a) = f(y,a)\}| = |U|$, then $Dens(x) = 1$.

In the universe, the more $Dens(x)$ is, if can be expressed in a graph, the more the number of objects around the x is, as well as the more possible x be a cluster center. So we select the object with the maximum average density as the first initial cluster center. For selection of the rest of initial cluster centers, we consider not only the distance between objects, but also the average density of the object. If the distance between the object and the already existing cluster centers is the only considered factor, it is possible that outlier is taken as a new cluster center. Similarly, if the density of the object is only taken into account, it is utmost possible that many cluster centers locate in the surrounding of one center. To avoid these potential problems, we combine the distance between objects with the density of the object together to measure the possibility of an object to be an initial cluster center. In the following, a new initialization method for categorical data is described as follows:

Input: $IS = (U, A, V, f)$ and k, where k is the number of cluster desired.
Output: *Centers*.
Step 1: $Centers = \emptyset$.
Step 2: For each $x_i \in U$, calculate the $Dens(x_i)$, $Centers = Centers \bigcup \{x_{i_1}\}$, where x_{i_1} satisfies $Dens(x_{i_1}) = max_{i=1}^{|U|}\{Dens(x_i)\}$, the first cluster center is selected.
Step 3: Find the second cluster center, $Centers = Centers \bigcup \{x_{i_2}\}$, where x_{i_2} satisfies $d(x_{i_2}, x_m) \times Dens(x_{i_2}) = max_{i=1}^{|U|}\{d(x_i, x_m) \times Dens(x_i) | x_m \in Centers\}$, goto Step 4.
Step 4: If $|Centers| \leqslant k$, then goto Step 5, otherwise goto Step 6.
Step 5: For any $x_i \in U$, $Centers = Centers \bigcup \{x_{i_3}\}$, where x_{i_3} satisfies $d(x_{i_3}, x_m) \times Dens(x_{i_3}) = max\{min_{x_m}\{d(x_i, x_m) \times Dens(x_i)\} | x_i \in U\}$, goto Step 4.
Step 6: End.

The time complexity of the proposed algorithm is composed of three parts. Firstly, we obtain the first initial cluster center, whose time complexity is $O(|U||A|)$. Secondly, computation of the second initial clusters center will take $O(|U||A|)$ steps. Finally, the rest of the initial clusters centers will take $O(|U||A|k^2)$ steps. Therefore, the whole time complexity of the proposed algorithm is $O(|U||A|k^2)$, which is linear with respect to the number of data objects.

4. Experimental analysis

In this section, in order to evaluate the efficiency of the initial cluster centers, we introduce an evaluation method (Yang, 1999) and some standard data sets are downloaded from the UCI Machine Learning Repository (UCI, 2006). All missing attribute values are treated as special values. Since there are no universally accepted methods for selecting initial cluster, we compare the results of the proposed method with random initialization method. Aiming at random initialization method, we carried out 100 runs of the k-modes algorithm and the fuzzy k-modes algorithm with different dissimilarity measures on these standard data sets, respectively. In each run, the same initial cluster centers randomly selected were used in the k-modes algorithm and the fuzzy k-modes algorithm with different dissimilarity measures. For the fuzzy k-modes algorithm we specified $\alpha = 1.1$ (Huang & Ng, 1999). Experimental results show that the proposed initialization method outperforms random initialization method.

4.1. Evaluation method

To evaluate the performance of clustering algorithms, an evaluation method is introduced (Yang, 1999). If 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 precision, recall and accuracy are defined as follows: $PR = \frac{\sum_{i=1}^{k}\left(\frac{a_i}{a_i+b_i}\right)}{k}$, $RE = \frac{\sum_{i=1}^{k}\left(\frac{a_i}{a_i+c_i}\right)}{k}$, $AC = \frac{\sum_{i=1}^{k} a_i}{|U|}$, respectively.

4.2. Data set

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

4.2.1. 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: D1, D2, D3, and D4. Except for D4, which has 17 instances, all the other diseases only have 10 instances each. We only selected 21 attributes in this experiments because the other attributes only have one category. The clustering results of the k-modes algorithm and the fuzzy k-modes algorithm on the soybean data are summarized in Tables 1 and 2, respectively.

4.2.2. 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 clustering results of the k-modes algorithm and the fuzzy k-modes algorithm on the zoo data are summarized in Tables 3 and 4, respectively.

4.2.3. Breast-cancer data

Breast-cancer data set consists of 699 data objects and 9 categorical attributes. It has two clusters Benign(458 data objects) and Malignant(241 data objects). The clustering results of the k-modes algorithm and the fuzzy k-modes algorithm on the breast-cancer data are summarized in Tables 5 and 6, respectively.

4.2.4. Mushroom data

Mushroom data set consists of 8124 data objects and 23 categorical attributes. Each object belongs to one of the two classes, edible(e) and poisonous(p). The clustering results of the k-modes

Table 6
The summary of clustering result of the fuzzy k-modes on the breast-cancer data.

The Fuzzy k-modes algorithm	Huang's dissimilarity measure		Ng's dissimilarity measure	
	Random	The proposed method	Random	The proposed method
AC	0.8031	0.9113	0.8259	0.8727
PR	0.8365	0.9292	0.8631	0.9110
RE	0.7199	0.8773	0.7528	0.8183
Iteration times	1.3500	2.0000	2.0000	2.0000

Table 7
The summary of clustering result of the *k*-modes on the mushroom data.

The *k*-modes algorithm	Huang's dissimilarity measure		Ng's dissimilarity measure	
	Random	The proposed method	Random	The proposed method
AC	0.7318	0.8754	0.7860	0.8922
PR	0.7520	0.9019	0.7969	0.9083
RE	0.7278	0.8709	0.7825	0.8887
Iteration times	3.2600	1.0000	3.5200	2.0000

Table 8
The summary of clustering result of the fuzzy *k*-modes on the mushroom data.

The Fuzzy *k*-modes algorithm	Huang's dissimilarity measure		Ng's dissimilarity measure	
	Random	The proposed method	Random	The proposed method
AC	0.7094	0.8754	0.7579	0.8872
PR	0.7231	0.9013	0.7672	0.9015
RE	0.7049	0.8709	0.7548	0.8839
Iteration times	2.9400	1	3.5200	2

algorithm and the fuzzy *k*-modes algorithm on the mushroom data are summarized in Tables 7 and 8, respectively.

From the above experiential results, for the *k*-modes algorithm and the fuzzy *k*-modes algorithm with different dissimilarity measures, one can see that the proposed method is superior to random initialization method with respect to *AC*, *PR*, and *RE*, respectively. In addition, we can also obtain the following conclusion that, for random initialization method, Ng's dissimilarity measure is better than Huang's dissimilarity measure in *k*-modes algorithm and fuzzy *k*-modes algorithm on these data set, respectively, and for the proposed method, Huang's dissimilarity measure is better than Ng's dissimilarity measure in *k*-modes algorithm and fuzzy *k*-modes algorithm except mushroom data set, respectively. Meanwhile, in most cases, the *k*-modes algorithm outperforms the fuzzy *k*-modes algorithm.

5. Conclusions

Categorical data are ubiquitous in real-world databases. The development of the *k*-modes type algorithm was motivated to solve this problem. However, these clustering algorithms need to be 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, the density of the object based on the frequency of attribute values has been defined. By taking into account the distance between the objects and the density of the object, a new initialization method for categorical data clustering has been proposed. Furthermore, the time complexity of the proposed algorithm has been analyzed. We tested the algorithm using four real world data sets from UCI Machine Learning Repository and experimental results have shown that the proposed method is superior to random initialization method in the *k*-modes algorithm and the fuzzy *k*-modes algorithm with different dissimilarity measures, respectively.

Acknowledgements

The authors are very grateful to the anonymous reviewers and editor. Their many helpful and constructive comments and suggestions helped us to significantly improve this work. This work was also supported by the National Natural Science Foundation of China (Nos. 60773133, 60573074, 60875040), the High Technology Research and Development Program of China (No. 2007AA01Z165), the National Key Basic Research and Development Program of China (973) (No. 2007CB311002), the Doctor Authorization Foundation of the Ministry of Education(No. 200801080006), the Natural Science Foundation of Shanxi(No. 2008011038), the Key Laboratory Open Foundation of Shanxi (Nos. 200603023, 2007031017) and the Technology Research Development Projects of Shanxi (No. 2007103).

References

Bradley, P. S., & Fayyad, U. M. (1998). Refining initial points for *K*-means clustering. In J. Sharlik (Ed.), *Proceedings of the 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, 315*(16), 972–976.

Duda, R. O., & Hart, P. E. (1973). *Pattern classification and scene analysis*. NY: John Wiley and Sons.

Fisher, D. (1987). Knowledge acquisition via incremental conceptual clustering. *Machine Learning, 2*, 139–172.

Fisher, D. (1996). Iterative optimization and simplification of hierarchical clusterings. *Journal of Artificial Intelligence Research, 4*, 147–179.

Gan, G., Wu, J., & Yang, Z. (2009). A genetic fuzzy *k*-modes algorithm for clustering categorical data. *Expert Systems with Application, 36*(2), 1615–1620.

Higgs, R. E., Bemis, K. G., Watson, I. A., & Wikel, J. H. (1997). Experimental designs for selecting molecules from large chemical databases. *Journal of Chemical Information and Computer Sciences, 37*, 861–870.

Huang, Z. X. (1997a). Clustering large datasets with mixed numeric and categorical values. In *Proceedings of the 1st Pacific Asia knowledge discovery and data mining conference* (pp. 21–34). Singapore: World Scientific.

Huang, Z. X. (1997b). A fast clustering algorithm to cluster very large categorical data sets in data mining. *Proceding SIGMOD Workshop Research Issues on Data Mining and Knowledge Discovery*, 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. K. (1999). A Fuzzy *k*-modes algorithm for clustering categorical data. *IEEE Transactions on Fuzzy Systems, 7*(4), 446–452.

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

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

Meila, M., & Heckerman, D. (1998). An experimental comparison of several clustering and initialization methods. In *Proceedings of the 14th conference on uncertainty in artificial intelligence* (pp. 386–395). San Francisco, CA: Morgan Kaufman.

Milligan, G. W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrica, 45*, 325–342.

Ng, M. K., Li, M. J., Huang, Z. X., & He, Z. Y. (2007). On the impact of dissimilarity measure in *k*-modes clustering algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*(3), 503–507.

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 comparison of four initalization methods for the *K*-means algorithm. *Pattern Recognition Letter, 20*, 1027–1040.

Snarey, M., Terrett, N. K., Willet, P., & Wilton, D. J. (1997). Comparison of algorithms for dissimilarity-based compound selection. *Journal of Molecular Graphics and Modelling, 15*, 372–385.

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 (2006). <http://www.ics.uci.edu/mlearn/MLRepository.html>.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of The American Statistical Association, 58*, 236–244.

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.

Zhang, W. X., Wu, W. Z., Liang, J. Y., & Li, D. Y. (2001). *Rough set theory and method*. Beijing, China: Science Press.