Expert Systems with Applications 38 (2011) 2381–2385

Contents lists available at ScienceDirect

Expert Systems with Applications

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

A data labeling method for clustering categorical data

Fuyuan Cao, Jiye Liang *

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

ARTICLE INFO

Keywords:
Data labeling
Categorical data
Rough membership function
Similarity measure

ABSTRACT

As the size of data growing at a rapid pace, clustering a very large data set inevitably incurs a time-consuming process. To improve the efficiency of clustering, sampling is usually used to scale down the size of data set. However, with sampling applied, how to allocate unlabeled objects into proper clusters is a very difficult problem. In this paper, based on the frequency of attribute values in a given cluster and the distributions of attribute values in different clusters, a novel similarity measure is proposed to allocate each unlabeled object into the corresponding appropriate cluster for clustering categorical data. Furthermore, a labeling algorithm for categorical data is presented, and its corresponding time complexity is analyzed as well. The effectiveness of the proposed algorithm is shown by the experiments on real-world data sets.

1. Introduction

The goal of clustering (Han & Kamber, 2001) is to partition a data set consisting of n points embedded in m-dimensional space into k distinct set of clusters, such that the data points within the same cluster are more similar to each other than to data points in other clusters according to some objective function that defines the similarity or dissimilarity among objects. Clustering has been studied in many research areas including pattern recognition, machine learning, statistical learning, data mining, text mining, and bioinformatics (Duda, Hart, & Stork, 2001; Hastie, Tibshirani, & Friedman, 2001; Jain & Dubes, 1988; Jain, Murty, & Flynn, 1999) and has been used in numerous application domains (Berkhin, 2002; Chen, Han, & Yu, 1996; Jain, Duin, & Mao, 2000; Xu & Wu, 2005).

As the size of data growing at a rapid pace, clustering a very large data set inevitably incurs a time-consuming process. Sampling has been recognized as an important technique to improve the efficiency of clustering. A typical approach to utilize sampling techniques on clustering is to randomly choose a small data set from the original data set, and the clustering algorithm is executed on the small sampled set. However, with sampling applied, those objects that are not sampled will not have their labels. For example, when we perform clustering for customers' segmentation with sampling techniques, a part of customers are sampled and grouped after clustering. Thus, the other customers that are not sampled will not obtain the cluster labels and do not belong to any segments. Without loss of generality, the goal of clustering is to allocate every object into an appropriate cluster. Therefore, an efficient method which is able to allocate the unclustered objects into appropriate cluster is very necessary.

In the numerical domain, there is a common solution to measure the similarity between an unclustered object and a cluster based on the distance between the unclustered object and the centroid of that cluster (Jain et al., 1999). Each unclustered object can be allocated to the cluster with the minimal distance. However, categorical attributes also prevalently exist in the real world. Minkowski metric, which is only to numeric data, becomes difficult to capture this notion for categorical attributes. Therefore, how to allocate unlabeled data object into proper clusters remains a challenging issue in the categorical domain.

Recently, assigning the unclustered categorical data objects to the clusters generated by using the sampled objects has attracted some attention. Similar to centroid (mean) of numerical attribute, in k-modes (Huang, 1998) algorithm, a cluster is represented by "mode", which is composed by the most frequent attribute value in each attribute domain in this cluster. Therefore, the similarity between an unclustered object and a cluster, namely the similarity between an unclustered object and the mode of a cluster, can be calculated by simple similarity matching, i.e., comparing two identical categorical values yields a difference of *one*, while comparing two distinct categorical values yields a difference of *zero*. The unclustered object is labeled to the cluster that obtained the maximal similarity. However, modes are not unique, and it is questionable using only one attribute value in each attribute domain to represent a cluster. ROCK (Guha, Rastogi, & Shim, 1999) uses the concept of a *link* to measure the similarity between categorical patterns. A measure $link(p_i,p_j)$ is defined as the number of common neighbors between two patterns p_i and p_j. The objective of the algorithm is to group together patterns that have more links. In

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

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.08.026

the final labeling phase of ROCK algorithm, clusters are represented by several representative objects, and each pattern x of the remaining patterns is assigned to the cluster l such that x has the maximum neighbors in lth cluster. Although ROCK provides high quality on the clustering results, the allocating procedure that measures the similarity between unclustered objects and each representative object is a very time-consuming process. In Chen, Chuang, and Chen (2008), a model named maximal resemblance data labeling (MARDL) was proposed to allocate each unclustered object into the corresponding appropriate cluster based on the novel categorical clustering representative. Although MARDL exhibits high execution efficiency and can achieve high intra-cluster similarity and low inter-cluster similarity, a user-specified parameter is essential to pruning algorithm and the NNIR tree constructing needs cost. The above-mentioned methods entirely adopted the representative of the cluster; however, it is very difficult to characterize and summary the clustering results by the representative of the cluster.

In this paper, a mechanism named rough membership function-based similarity (abbreviated as RMFS) is to allocate each unclustered categorical data object into the corresponding proper cluster. The allocating process is referred to as data labeling: to give each unclustered object a cluster label. For simplicity, we call the unclustered object as unlabeled object in the sequel. These unlabeled objects will be allocated into clusters by two phases, namely, the cluster analysis phase and the data labeling phase, which are briefly described below. In the cluster analysis phase, we can generate clustering result on sampled set by choosing the corresponding categorical clustering algorithms, such as k-modes algorithm (Huang, 1998), fuzzy k-modes algorithm (Huang & Ng, 1999), fuzzy k-modes algorithm with fuzzy centroid (Kim, Lee, & Lee, 2004), and so on. In the data labeling phase, according to RMFS similarity measure, each unlabeled object is allocated into the cluster which possesses the maximal similarity. Furthermore, a labeling algorithm for categorical data is presented, and its corresponding time complexity is analyzed as well. The effectiveness of the proposed algorithm is validated by the experiments on real-world data sets.

The outline of the rest of this paper is as follows. In Section 2, the similarity between an unlabeled object and a cluster is defined. In Section 3, a labeling algorithm for categorical data is proposed, and its corresponding time complexity is analyzed. Section 4 illustrates the performance study on real data sets. Section 5 concludes the paper.

2. Similarity measure between a cluster and an unlabeled object

In this section, for convenience, we first give the definition of a categorical information system. In the following, we review some basic concepts of rough set theory, such as the indiscernibility relation, rough membership function, and so on. Then, a novel similarity between an unlabeled object and a cluster is defined by considering the frequency of attribute values in a given cluster and the distributions of attribute values in different clusters.

In general, we assume the set of objects to be clustered is stored in a table, where each row (tuple) represents the fact about an object. A data table is also called an information system. Data in the real world are prevalently described by categorical attributes. More formally, a categorical information system can be defined as a quadruple $IS = (U, A, V, f)$, where U is the nonempty set of objects, called the universe, A is the nonempty set of attributes, V is the union of all attribute domains, i.e., $V = \bigcup_{a\in A} V_a$, where V_a is the domain of attribute a and it is finite and unordered.

f: $U \times A \to V$ – a mapping called an information function such that for any $x \in U$ and $a \in A$, $f(x,a) \in V_a$.

Rough set theory introduced by Pawlak (1982) is a kind of symbolic machine learning technology for categorical information systems with uncertainty information (Liang & Li, 2005; Liang, Wang, & Qian, 2009). In recent years, rough set theory has attracted close attention in some clustering literatures. Parmar, Wu, and Blackhurst (2007) proposed a new algorithm MMR (Min-Min-Roughness) for clustering categorical data based on rough set theory, which has the ability to handle the uncertainty in the clustering process. By the notion of rough membership function in rough set theory, Jiang, Sui, and Cao (2008) defined the rough outlier factor for outlier detection. Chen and Wang (2006) presented an improved clustering algorithm based on rough set and Shannon's Entropy theory. Based on the neighborhood-based rough set model, an initialization method for the k-means algorithm was presented (Cao, Liang, & Jiang, 2009).

For ease of presentations, we first review some basic concepts of rough set theory.

Definition 1. Let $IS = (U, A, V, f)$ be a categorical information system, for any attribute subset $P \subseteq A$, a binary relation $IND(P)$, called indiscernibility relation, is defined as

$$IND(P) = \{(x,y) \in U \times U | \forall a \in P, f(x,a) = f(y,a)\}.$$

It is obvious that $IND(P)$ is an equivalence relation on U and $IND(P) = \bigcap_{a\in P} IND(\{a\})$.

Given $P \subseteq A$, the relation $IND(P)$ induces a partition of U, denoted by $U/IND(P) = \{[x]_P^U | x \in U\}$, where $[x]_P^U$ denotes the equivalence class determined by x with respect to P, i.e., $[x]_P^U = \{y \in U | (x,y) \in IND(P)\}$.

Definition 2. Let $IS = (U, A, V, f)$ be a categorical information system, $P \subseteq A$ and $X \subseteq U$. The rough membership function $\mu_{U,X}^P : U \to [0,1]$ is defined as

$$\mu_{U,X}^P(x) = \frac{|[x]_P^U \bigcap X|}{|[x]_P^U|}.$$

The rough membership function quantifies the degree of relative overlap between the set X and the equivalence class $[x]_P^U$ to which x belongs.

In classical set theory, an element either belongs to a set or not. The corresponding membership function is the characteristic function for the set, i.e., the function takes values 1 and 0, respectively. However, the rough membership function takes values between 0 and 1.

Example 1. Consider the data set in Table 1.

This is a categorical information system, where $U = \{x_1, x_2, \ldots, x_{19}\}$ and $A = \{A_1, A_2, A_3\}$. Let $X = \{x_1, x_2, \ldots, x_{10}\}$ and $P = \{A_1, A_2\}$.

By Definition 2, it is easy to obtain that

$$\mu_{U,X}^P(x_{17}) = \frac{|[x_{17}]_P^U \bigcap X|}{|[x_{17}]_P|} = \frac{\{x_7, x_{11}, x_{13}, x_{17}\} \bigcap \{x_1, x_2, \ldots, x_{10}\}}{\{x_7, x_{11}, x_{13}, x_{17}\}} = \frac{1}{4}.$$

In order to measure the similarity between an unlabeled object and a cluster, an improved definition of rough membership function is given.

Definition 3. Let $IS = (U, A, V, f)$ be a categorical information system, $P \subseteq A$, $U = S \cup Q$ and $S \cap Q = \emptyset$. For any $x \in Q$ and $X \subseteq S$, the rough membership function $\mu_{Q,X}^P : Q \to [0,1]$ is defined as

$$\mu_{Q,X}^P(x) = \begin{cases} \frac{|[x]_P^S \bigcap X|}{|[x]_P^S|}, & \text{if } [x]_P^S \neq \emptyset, \\ 0, & \text{otherwise}, \end{cases}$$

where

Table 1
An example data set.

Objects	A_1	A_2	A_3
x_1	a	m	c
x_2	b	m	b
x_3	c	f	c
x_4	a	m	a
x_5	a	m	c
x_6	c	f	a
x_7	c	m	a
x_8	c	f	c
x_9	a	f	b
x_{10}	b	m	a
x_{11}	c	m	c
x_{12}	c	f	b
x_{13}	c	m	b
x_{14}	b	m	c
x_{15}	a	f	a
x_{16}	a	m	c
x_{17}	c	m	a
x_{18}	b	f	b
x_{19}	a	f	c

$$[x]_P^S = \{u \in S | \forall a \in P, f(u,a) = f(x,a)\}.$$

In Definition 3, the domain of the rough membership function is a subset Q of U, not the universe U. Moreover, we only consider the equivalence class of x on set S with respect to attribute set P.

For the rough membership function $\mu_{Q,X}^P$ of Definition 3, it is easy to prove the following properties.

Property 1. *If* $[x]_P^S \subseteq X$, *then* $\mu_{Q,X}^P(x) = 1$;

Property 2. *If* $[x]_P^S \bigcap X = \emptyset$ *or* $[x]_P^S = \emptyset$, *then* $\mu_{Q,X}^P(x) = 0$;

Property 3. *For any* $a \in P$, *if* $x, y \in Q$ *and* $f(x, a) = f(y, a)$, *then* $\mu_{Q,X}^P(x) = \mu_{Q,X}^P(y)$;

Property 4. *If* $X, Y \subseteq S$ *and* $X \subseteq Y$, *for any* $x \in Q$, *then* $\mu_{Q,X}^P(x) \leqslant \mu_{Q,Y}^P(x)$.

Example 2 (*Continued from Example 1*). Let $S = \{x_1, x_2, \ldots, x_{15}\}$, $Q = U - S$, $X = \{x_1, x_2, \ldots, x_{10}\}$ and $P = \{A_1, A_2\}$. Then,

$$\mu_{S,X}^P(x_{17}) = \frac{|[x_{17}]_P^S \bigcap X|}{|[x_{17}]_P^S|} = \frac{\{x_7, x_{11}, x_{13}\} \bigcap \{x_1, x_2, \ldots, x_{10}\}}{\{x_7, x_{11}, x_{13}\}} = \frac{1}{3}.$$

Definition 3 quantifies the degree of relative overlap between the subset X of S and the equivalence class $[x]_P^S$ to which x belongs. As the number of the attribute grows, the number of equivalence class of the unlabeled object with respect to the given attribute set decreases, which usually results in $\mu_{S,X}^P(x) = 0$. To overcome these drawbacks, a novel similarity between an unlabeled object and a cluster is defined as follows.

Definition 4. Let $IS = (U, A, V, f)$ be a categorical information system, $P \subseteq A$, $U = S \cup Q$ and $S \cap Q = \emptyset$. Suppose that a prior clustering result $S = \{c_1, c_2, \ldots, c_k\}$ is given, where c_i, $1 \leqslant i \leqslant k$, is the ith cluster. For any $x \in Q$, the similarity between an unlabeled object x and a cluster c_i with respect to P is defined as

$$Sim_P(x, c_i) = \sum_{a \in P} \sqrt{w_a} \times m'_a,$$

where

$$w_a = \frac{|[x]_{\{a\}}^S \bigcap c_i|}{|[x]_{\{a\}}^S|},$$

$$m'_a = \frac{|\{u | f(u,a) = f(x,a), u \in c_i\}|}{|c_i|}$$

and $|c_i|$ is the number of objects in the ith cluster. w_a considers the distribution of attribute value $f(x,a)$ between clusters. m'_a characterizes the importance of the attribute value $f(x,a)$ in the cluster c_i with respect to attribute a. Hence, $Sim_P(x, c_i)$ considers both the intra-cluster similarity and the inter-cluster similarity.

Example 3 (*Continued from Example 1*). Let $S = \{c_1, c_2, c_3\}$, where $c_1 = \{x_1, x_2, \ldots, x_5\}$, $c_2 = \{x_6, x_7, \ldots, x_{10}\}$, $c_3 = \{x_{11}, x_{12}, \ldots, x_{15}\}$, $P = A = \{A_1, A_2, A_3\}$.

According to Definition 4, it is clear that

$$Sim_P(x_{17}, c_1) = \sum_{a \in P} \sqrt{w_a} \times m'_a = \sqrt{\frac{1}{7}} \times \frac{1}{5} + \sqrt{\frac{4}{9}} \times \frac{4}{5} + \sqrt{\frac{1}{5}} \times \frac{1}{5} = 0.6984,$$

$$Sim_P(x_{17}, c_2) = \sum_{a \in P} \sqrt{w_a} \times m'_a = \sqrt{\frac{3}{7}} \times \frac{3}{5} + \sqrt{\frac{2}{9}} \times \frac{2}{5} + \sqrt{\frac{3}{5}} \times \frac{3}{5} = 1.0461,$$

and

$$Sim_P(x_{17}, c_3) = \sum_{a \in P} \sqrt{w_a} \times m'_a = \sqrt{\frac{3}{7}} \times \frac{3}{5} + \sqrt{\frac{3}{9}} \times \frac{3}{5} + \sqrt{\frac{1}{5}} \times \frac{1}{5} = 0.8286.$$

Similarly, we have that

$$Sim_P(x_{16}, c_1) = 1.4224, \quad Sim_P(x_{16}, c_2) = 0.3597,$$
$$Sim_P(x_{16}, c_3) = 0.6668,$$
$$Sim_P(x_{18}, c_1) = 0.2971, \quad Sim_P(x_{18}, c_2) = 0.6397,$$
$$Sim_P(x_{18}, c_3) = 0.6293$$

and

$$Sim_P(x_{19}, c_1) = 0.9707, \quad Sim_P(x_{19}, c_2) = 0.5954,$$
$$Sim_P(x_{19}, c_3) = 0.5513.$$

Obviously, x_{17} and x_{18} can be allocated the cluster c_2. x_{16} and x_{19} are labeled to the cluster c_1.

Note that after executing the data labeling phase, the unlabeled objects obtain a cluster label but is not really added to the cluster.

3. Data labeling algorithm based on RMFS

The goal of clustering is to allocate every data object into an appropriate cluster. In the cluster analysis phase, we can generate clustering result by choosing the corresponding categorical clustering algorithm, such as the k-modes algorithm on sampled data set. In the data labeling phase, each unlabeled object is given a label of appropriate cluster according to RMFS. The pseudocode of labeling unlabeled categorical data algorithm based on RMFS is described in Table 2.

The runtime complexity of cluster analysis phase is dependent on the using categorical clustering algorithm. The runtime complexity of data labeling phase can be analyzed as follows. The runtime complexity for computing the similarity between arbitrary unlabeled object and a cluster is $O(|S||P|)$. Therefore, the whole computational cost of the proposed algorithm is $O(|S||P||Q|k)$. Based on the above analysis, the time complexity on the data labeling

Table 2
Data labeling algorithm based on RMFS.

```
1 Input: IS = (S ⋃ Q, A, V, f), where S is a sampled data, Q is an unlabeled data
    set,
2 k is the number of clusters;
3 Generate a partition S = {c1,c2,...,ck} of S with respect to A by calling the
4 corresponding categorical clustering algorithm;
5 For i = 1 to |Q|
6   For j = 1 to k
7     calculate the similarity between the ith object and the jth cluster
8     according to Definition 4, and the ith object is labeled to the cluster
9     that obtained the maximum similarity;
10  End;
11 End;
12 Output: each object of Q is labeled to the cluster that obtained
13 the maximum similarity.
```

phase is linear with respect to the number of the objects in the unlabeled data set Q.

4. Experimental analysis

In this section, we demonstrate the effectiveness of the proposed algorithm. In Section 4.1, the test environment and the data set used are described. Section 4.2 presents the accuracy of the proposed algorithm on the real data sets.

4.1. Simulation environment and data sets

All of our experiments are conducted on a PC with Intel Pentium D (2.8 G) processor, 1 GB memory, and Windows XP SP3 professional operating system installed. In all the experiments, the random sample technique is used for data sampling, and the k-modes algorithms with the two different measure are chosen to do clustering on the sampled data set, respectively. One real data set, Mushroom data set (UCI Machine Learning Repository, 2009) is used in our study to demonstrate the effectiveness of the RMFS. In the following, the real data set which is utilized in the experiments is described.

There are 8124 objects described by 22 categorical attributes in Mushroom data set. Each object describes the physical characteristics of a single mushroom. Each object belongs to one of two classes: edible (e) and poisonous (p). The mushroom data set contains 2480 objects with missing attribute values. In order to test the stability of the proposed algorithm, we discuss two cases in the experiments. One case is that missing attribute values are replaced with a special value, the other case is the objects with missing attribute values are removed.

4.2. Evaluation on accuracy

In this evaluation, the accuracy of the proposed algorithm is compared against the results of clustering entire data set. The accuracy of clustering and data labeling are calculated by the following equations:

$$AC_U = \frac{\sum_{i=1}^{k} a_i}{|U|},$$

and

$$AC_Q = \frac{\sum_{i=1}^{k} b_i}{|Q|},$$

respectively, where k is the number of clusters of the data, a_i is the number of objects that are correctly assigned to the class c_i on the U, b_i is the number of objects that are correctly assigned to the class c_i on the unlabeled set Q.

Table 3
The clustering accuracy of the k-modes algorithm with the two dissimilarity measures on entire data set.

Data set	Huang's measure	Ng's measure
Mushroom (removed)	0.7829	0.7868
Mushroom (replaced)	0.7218	0.7836

Table 4
The accuracy comparison of different size: using the k-modes algorithm with Huang's measure on the sampled data set.

Size (%)	Mushroom (removed)	Mushroom (replaced)
1	0.7808	0.7412
3	0.7798	0.7497
5	0.8173	0.7676
7	0.7855	0.7653
9	0.7960	0.7618

Table 5
The accuracy comparison of different size: using the k-modes algorithm with Ng's measure on sampled data set.

Size (%)	Mushroom (removed)	Mushroom (replaced)
1	0.7755	0.7517
3	0.7369	0.7929
5	0.7872	0.8086
7	0.7921	0.7920
9	0.7965	0.8165

First, we carried out 50 runs of the k-modes algorithm with the two different dissimilarity measures (Huang, 1998; Ng, Li, Huang, & He, 2007) on Mushroom data sets, respectively. In each run, the same initial cluster centers were used for the two different dissimilarity measures. The experimental results are summarized in Table 3. Each value in Table 3 is the average of 50 times experiments.

From Table 3, we can find that Ng's measure is superior to Huang's measure.

Furthermore, the k-modes algorithms with the two different measures are chosen to do clustering on sampled data set, respectively, and the clustering accuracy on Mushroom data set are shown in Tables 4 and 5, respectively.

Comparing Table 3 with Tables 4 and 5, we can find that the accuracy of RMFS with different size mostly outperforms that of the entire data set on using Huang's measure or Ng's measure. The situation is bad when size is 1% and 3%. This result can be explained by the reason that the number of objects of sampled data may be too small. In addition, as the k-modes algorithm is dependent on the initial centers, the clustering results obtained from the sample data set may not be similar to that obtained from the original data set. Therefore, the data labeling results are influenced directly by the clustering result obtained from the sampled data set.

5. Conclusions

Data labeling techniques have become an important issue in machine learning. In the categorical domain, the problem of how to allocate the unlabeled objects into appropriate clusters has not been fully explored in the previous works. In this paper, based on the rough membership function and the frequency of attribute values, a new similarity measure for allocating the unlabeled objects into appropriate clusters has been defined. A distinct characteristic of the new similarity measure is to characterize the

distribution of the attribute values in the different clusters and the frequency of the attribute values in given cluster, which consider both the intra-cluster similarity and the inter-cluster similarity. Based on the proposed similarity measure, a data labeling algorithm has been presented for clustering categorical data, and its corresponding time complexity has been analyzed as well. The results of comparative experiments on Mushroom data sets from UCI have shown the effectiveness of the new similarity measure.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Nos. 60773133, 70971080 and 71031006), the National Key Basic Research and Development Program of China (973) (No. 2007CB311002), the Natural Science Foundation of Shanxi (Nos. 2008011038 and 2010021016-2), and the Technology Research Development Projects of Shanxi (No. 2007103).

References

Berkhin, P. (2002). *Survey of clustering data mining techniques*. Technical Report, Accrue Software, San Jose, CA.

Cao, F. Y., Liang, J. Y., & Jiang, G. (2009). An initialization method for the *k*-means algorithm using neighborhood model. *Computers and Mathematics with Applications, 58*(3), 474–483.

Chen, H. L., Chuang, K. T., & Chen, M. S. (2008). On data labeling for clustering categorical data. *IEEE Transactions on Knowledge and Data Engineering, 20*(11), 1458–1471.

Chen, M. S., Han, J., & Yu, P. S. (1996). Data Mining: An overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineer, 8*(6), 866–883.

Chen, C. B., & Wang, L. Y. (2006). Rough set-based clustering with refinement using Shannon's entropy theory. *Computer and Mathematics with Application, 52*(10–11), 1563–1576.

Duda, R., Hart, P., & Stork, D. (2001). *Pattern classification*. John Wiley & Sons.

Guha, S., Rastogi, R., & Shim, K. (1999). Rock: A robust clustering algorithm for categorical attributes. In *Proceedings of the IEEE international conference on data engineering, Sydney, Australia* (pp. 512–521).

Han, J., & Kamber, M. (2001). *Data mining concepts and techniques*. Morgan Kaufmann: San Francisco.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference and prediction*. Springer.

Huang, Z. X. (1998). Extensions to the *k*-means algorithm for clustering large data sets with categorical values. *Data Mining and 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.

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

Jain, A. K., Duin, R. P., & Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(1), 4–37.

Jain, A., Murty, M., & Flynn, P. (1999). Data clustering: A review. *ACM Computing Surveys, 31*(3), 264–323.

Jiang, F., Sui, Y. F., & Cao, C. G. (2008). A rough set approach to outlier detection. *International Journal of General Systems, 37*(5), 519–536.

Kim, D. W., Lee, K. H., & Lee, D. (2004). Fuzzy clustering of categorical data using fuzzy centroids. *Pattern Recognition Letters, 25*(11), 1263–1271.

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

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

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.

Parmar, D., Wu, T., & Blackhurst, J. (2007). MMR: An algorithm for clustering data using rough set theory. *Data & Knowledge Engineering, 63*(3), 893–897.

Pawlak, Z. (1982). Rough sets. *International Journal of Computer and Information Sciences, 11*, 341–356.

UCI Machine Learning Repository. <http://www.ics.uci.edu/mlearn/MLRepository.html>, 2009.

Xu, R., & Wu, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks, 16*(3), 645–678.