International Journal of Database Theory and Application
Vol.6, No.5 (2013), pp.95-104
http://dx.doi.org/10.14257/ijdta.2013.6.5.09

Clustering Algorithm for Incomplete Data Sets with Mixed Numeric and Categorical Attributes

Wu Sen, Chen Hong and Feng Xiaodong

Dongling School of Economics and Management
University of Science and Technology Beijing
Beijing, 100083, P.R.China
wusen@manage.ustb.edu.cn

Abstract

The traditional k-prototypes algorithm is well versed in clustering data with mixed numeric and categorical attributes, while it is limited to complete data. In order to handle incomplete data set with missing values, an improved k-prototypes algorithm is proposed in this paper, which employs a new dissimilarity measure for incomplete data set with mixed numeric and categorical attributes and a new approach to select k objects as the initial prototypes based on the nearest neighbors. The improved k-prototypes algorithm can not only cluster incomplete data with no need to impute the missing values, but also avoid randomness in choosing initial prototypes. To illustrate the accuracy of the established algorithm, traditional k-prototypes algorithm and k-prototypes employing the new dissimilarity measure are compared to the improved k-prototypes algorithm by using data from UCI machine learning repository. The experimental results show that the improved k-prototypes algorithm is superior to the other two algorithms with higher clustering accuracy.

Keywords: *Mixed numeric and categorical attributes, k-prototypes algorithm, Initial k prototypes, Missing value imputation*

1. Introduction

Clustering has been applied in various areas in many years, for instance, the trend analysis of financial data, customer segmentation, automatic detection of pattern recognition, and so on. Undoubtedly, the capability of dealing with incomplete data sets with mixed numeric and categorical attributes is rather important for clustering algorithms because of extensively existence of incomplete data in real-world databases.

K-means algorithm is well-known to cluster numeric data. It uses mean as the center of a cluster. Since means merely exist in numeric attributes, Huang [1] proposes k-modes algorithm which replaces means of clusters with modes and applies a frequency based approach to update modes in the clustering process to extend k-means [2] in order to use a simple matching dissimilarity measure to process categorical attributes. However, the k-modes algorithm is still unable to handle the data with mixed values. Therefore, Huang [3] puts forward a k-prototypes algorithm which integrates k-means and k-modes to cluster mixed data. The k-prototypes algorithm redefines a dissimilarity measure that takes into account both numeric and categorical attributes, aiming at updating the cluster prototypes for ease of optimizing objective function.

While the k-prototypes algorithm still has some limitations in the following three aspects: clusters initialization, dissimilarity computation and isolated points. Scholars have been seeking for solutions to solve these problems from different points of view.

ISSN: 2005-4270 IJDTA

Firstly, selecting *k* objects as initial cluster centers affects clustering results, so clusters initialization optimization is necessary. Inappropriate choice of *k* initial centers may increase the number of iterations and time consumption in the clustering process [4]. Taking these negative effects into consideration, some researchers dedicate to optimize this algorithm. Zhao Lijiang [5] proposes to group all *n* data objects into *n*/*k* groups and randomly select an object from each group as the initial center respectively to get the mean and high-frequency values for numeric and categorical attributes.

Moreover, dissimilarity computation in categorical attributes needs to be improved. In small data sets, the traditional dissimilarity computation is able to aptly reflect difference between objects. But objects may be added to the cluster randomly if the dissimilarity between objects is same. So Wu Mengshu [6] refers to the idea of projection, pursuing to find the optimal projection direction, and then cluster the high-dimensional objects, which can effectively avoid the impact of dimensional disaster.

Finally, clustering results are sensitive to isolated points. Focused on isolated points [7], one solution is to employ object closest to the center position in the cluster instead of mean values in the cluster.

Nowadays, many researchers concentrate on how to cluster effectively and propose some efficient algorithms including five kinds of traditional clustering algorithms and many new clustering algorithms in six aspects. Thus clustering category [8-12] is shown in Table 1.

Table 1. Clustering Category

Clustering algorithms
a) Traditional Clustering algorithms
1) Partitioning：K-means, PAM, CLARA, K-modes, EM, CLARANS, ISODATA. 2) Hierarchical：BIRCH, CURE, Chameleon. 3) Grid：STING, Wave cluster, CLIQUE. 4) Density：DBSCAN, OPTICS, DENCLUE. 5) Model：COBWEB, CLASSIT, LVQ, SOM.
b) New Clustering algorithms
1) Ownership relation：Granular, Uncertainty, Spherical Shell, Entropy. 2) Data preprocessing：Kernel, Concept. 3) Similarity： Spectral, Affinity Propagation, Ontology, Hybrid data, Dual Distance. 4) Update strategy：Data Stream, PSO. 5) High dimensional：Projection Pursuit, Subspace. 6) Integration with other science：Quantum, Clustering Ensemble, Random Walk.

While dealing with incomplete data containing missing values is always challenging. It brings plenty of difficulties in clustering. Simple techniques to handle missing data (such as deletion of incomplete observations and the missing indicator method) produce biased

clustering results. Hence, a large number of scholars pour their energy into research of imputation.

Imputation is a process which imputes known or empirical value in missing data through some particular algorithms, and it has become a relatively popular approach to get complete data. Imputation principally employs three methods, namely imputations based on statistics, rough set theory and data mining techniques such as classification and clustering. Primarily, the first approach on the basis of statistics depends on the distribution of remained objects' values of the attributes, and it usually applies to numeric attributes. Chatzis Sotirios [13] puts forward c-mean substitution method which replacing missing data with mean values. The second method based on rough set theory utilizes the tolerance relation between objects which handles only categorical attributes. Wang Lei and Ma Weimin [14, 15] propose a rough set theory which makes the decision rules as high support degree as possible. The third approach exploits some related data mining techniques such as classification and clustering. Wang Fengmei [16] introduces that k-nearest neighbors imputation calculates the Euclidean distance between each object with missing values and other complete objects, and then selects *k* objects with the smallest distance to the incomplete object in order to estimate missing values with weighted average values. Xu Fang [17] proposes a modified shuffled frog leaping clustering method based on k-means imputation which partitions data into complete objects and incomplete objects, and then assigns the incomplete objects to the closest complete objects and fills in the missing values with cluster center.

Many of these imputation approaches, however, have their limitations as well. On one hand, only numeric attributes or categorical attributes could be dealt with by these approaches and some approaches simply fill the missing value with mean imputation [18], in this case more errors in the process of data mining might be caused. On the other hand, non-missing values of incomplete objects are ignored by imputation approaches above and these approaches merely concentrate on the observed values of complete objects.

Therefore, owing to these problems, our main focus in this paper is to study k-prototypes algorithm which has apriority on managing mixed data. An improved k-prototypes algorithm is proposed with a new approach to select initial *k* centers based on nearest neighbors and a new dissimilarity measure for incomplete data sets with mixed numeric and categorical attributes. Choosing *k* initial centers based on the idea of nearest neighbors decreases the randomness of traditional k-prototypes. The new dissimilarity measure has a comprehensive consideration over the incomplete objects information. After clustering has been carried out, the missing values can also be imputed on the basis of clustering results. According to the empirical experiments results, the improved k-prototypes algorithm produces higher clustering accuracy.

2. Materials and Methods

2.1. Data Sets

We use Thyroid disease data from UCI machine learning repository to assess the performance of the improved k-prototypes algorithm, which is based on dissimilarity measure for incomplete data set with mixed numeric and categorical attributes and an improved selection of initial *k* prototypes based on nearest neighbors. Thyroid disease data set is collected by New South Wales Institute, including 3428 records with each being described by 15 categorical attributes and 6 numeric attributes.

2.2. Problem Description

An incomplete data set $S = \{U, A, V, f\}$, where $U = \{x_1, x_2, \cdots, x_n\}$, $A = C \sqcup N = \{a_k | k = 1,2, \cdots, m\} \sqcup \{a_l | l = m+1, m+2, \cdots, m+q\}$. The number of objects is n; the number of attributes is $m+q$; C is the data set of categorical attributes; N is the data set of numeric attributes; V is the set of all values. f represents the function $U \times A \to V$. In this paper, we set the missing values by "*".

2.3. Incomplete Set Mixed Dissimilarity (*ISMD*)

Given the incomplete system $S = \{U, A, V, f\}$, X is a subset of U; x_i and x_j are two objects in X and $|X|$ is the number of objects in X. Here incomplete set mixed dissimilarity (*ISMD*) of X combines $ISMDC(x_i, x_j)$ and $ISMDN(x_i, x_j)$ with a certain weight is defined as:

$$ISMD(x_i, x_j) = \frac{w_c \times ISMDC(x_i, x_j) + w_n \times ISMDN(x_i, x_j)}{w_c + w_n} \quad (1)$$

Where w_c and w_n respectively represent the weight of categorical attributes and numeric attributes. Usually, we choose the number of categorical attributes m and the number of numeric attributes q as the weight correspondingly.

$ISMDC(x_i, x_j)$ represents the categorical attributes dissimilarity degree. According to the rough set theory [19], it can be defined as:

$$\delta_k(x_i, x_j) = \begin{cases} 1 & a_k(x_i) \neq a_k(x_j) \wedge a_k(x_i) \neq "*" \wedge a_k(x_j) \neq "*" \\ 0 & a_k(x_i) = a_k(x_j) \vee a_k(x_i) = "*" \vee a_k(x_j) = "*" \end{cases} \quad (2)$$

$$ISMDC(x_i, x_j) = \frac{\sum_{k=1}^{m} \delta_k(x_i, x_j)}{m - \sum_{k=1}^{m} \delta_k(x_i, x_j)} \quad (3)$$

Where $\delta_k(x_i, x_j)$ represents distance between x_i and x_j in categorical attribute k. In condition that the distance formula is related to the number of categorical attributes, we obtain incomplete categorical attributes dissimilarity through normalizing as above formula (3).

$ISMDN(x_i, x_j)$ represents the numeric attributes dissimilarity degree. Depending on the distance of minimum-maximum standardization and Minkowski distance, the formula is as follows:

$$d_l(x_i, x_j) = \begin{cases} \frac{|a_l(x_i) - a_l(x_j)|}{Max_l - Min_l} & a_l(x_i) \neq "*" \wedge a_l(x_j) \neq "*" \\ 0 & a_l(x_i) = "*" \vee a_l(x_j) = "*" \end{cases} \quad (4)$$

$$ISMDN(x_i, x_j) = \frac{(\sum_{l=m+1}^{m+q} d_l(x_1, x_2)^2)^{\frac{1}{2}}}{q - (\sum_{l=m+1}^{m+q} d_l(x_1, x_2)^2)^{\frac{1}{2}}} \quad (5)$$

Where Max_l and Min_l respectively represent the maximum and minimum value of attribute a_l in X. Meanwhile, $d_l(x_i, x_j)$ is the standardized numeric distance and $ISMDN(x_i, x_j)$ is calculated through normalizing as above formula (5).

In particular, the incomplete set mixed dissimilarity (*ISMD*) can deal with incomplete data sets with mixed numeric and categorical attributes. Furthermore, with no need for imputing mean values in advance, *ISMD* measures dissimilarity of data objects with missing values directly and this decreases the clustering errors indeed.

2.4. Improved Selection of Initial *k* Centers based on Nearest Neighbors

Selecting initial *k* centers is aimed at choosing the most representative objects having lots of nearest neighbors. This paper improves the selection process of initial *k* cluster centers based on the nearest neighbors method. Now, consider the basic idea of this selection strategy as follows.

Input: Pre-set threshold θ, initial empty center set P and an incomplete data set $S = \{U, A, V, f\}$;

Output: Initial center set *P'*;

Steps:

Step1: Initialization. Set *P* as an empty set.

Step2: Calculate the dissimilarity between each two objects based on formula (1).

Step3: Regard each two objects' dissimilarity which is less than Pre-set threshold θ as neighbors.

Step4: Select the object who owns the largest number of neighbors as the first initial center in *P*.

Step5: Refresh Pre-set threshold θ and delete the initial center with its neighbors in the data set *S*. In addition, the number of changing threshold θ needs to be *k-1* times in sure to get *k* initial centers.

Step6: Repeat step 3 to step 5 until the original data set *S* is ϕ and gain the initial *k* centers set *P'*.

Therefore, the obtained set *P'* is used as the initial *k* centers of k-prototypes algorithm, which apparently avoids the randomness of initial *k* centers selection.

2.5. Improved k-prototypes Algorithm for Incomplete Data with Mixed Attributes

Based on the analysis above, the main steps of the improved k-prototypes algorithm for incomplete data with mixed attributes are as follows:

Input: Clustering number *k* and an incomplete data set $S = \{U, A, V, f\}$;

Output: *k* clusters and data set *S'* after imputation;

Steps:

Step1: Initialization *k* prototypes. Select the initial *k* centers as prototypes $P^{(0)} = \{P_1, P_1, \cdots, P_k\}$ based on the nearest neighbors method.

Step2: Allocation. Search the minimum incomplete set mixed dissimilarity between object X_i and prototype P_l.Through calculating the dissimilarity $ISMD(x_i, x_j)$ between each object and each prototype, assign the object into the cluster with the nearest prototype until all objects are put into corresponding clusters.

Step3: Re-allocation. Calculate new *k* cluster prototypes and reassign each object. For each cluster, recalculate the new *k* cluster prototypes based on mean value of numeric attributes and high frequency value of categorical attributes. Then figure out the incomplete set mixed dissimilarity $ISMD(x_i, x_j)$ between each object and each new prototype. If the closest

prototype for the object is not the current one, redistribute the object into the cluster with the new nearest prototype.

Step4: Repeat step3 until the composition of *k* cluster prototypes don't change. And then turn to step 5.

Step5: Get *k* clusters and impute the missing values. Fill in the missing values with mean value M_i of its attribute in numeric data set and mode value $Mode_i$ in categorical data set. The formulas are as follows:

$$I_l = \{x_i \in X | a_l(x_i) = " * "\} \tag{6}$$

$$M_i = \begin{cases} \frac{1}{|X|-|I_l|}\sum_{x_i \notin I_l} a_l(x_i) & |X| > |I_l| \\ +\propto & |X| = |I_l| \end{cases} \tag{7}$$

Where $a_l(x_i)$ represents the value of object x_i on numeric attribute a_l.

$$Mode_i = \begin{cases} max(|a_k(x_i)|) & a_k(x_i) \neq " * " \\ 0 & a_k(x_i) = " * " \end{cases} \tag{8}$$

Where $a_k(x_i)$ represents the value of object x_i on categorical attribute a_k.

3. Numerical Results

In the experimental stage, we set the Pre-set threshold θ is 0.2 with the amplification of 0.1. And then set the missing rate γ of data set from the ratio of 5% to 20% with 5% interval. For each missing ratio in data set, we complete clustering and imputation respectively by traditional k-prototypes, k-prototypes employing the new dissimilarity measure *ISMD* (Abbreviated as k-prototypes+*ISMD*) and improved k-prototypes. We evaluate the performance of the algorithms by using three metrics: *RC*, *RMSSE* and γ given below.

3.1. Evaluation Indexes

(1) Accurate Rate in Categorical Attributes after Imputation *RC*

For the categorical attributes, accurate rate in categorical attributes after imputation *RC* is the ratio of average number of correct imputed categorical values per run and the initial number of the missing values. The formula is as follows:

$$\mathrm{RC} = \mathrm{C}/(\gamma * \mathrm{m} * \mathrm{n}) \tag{9}$$

C represents the number of correct imputed categorical values. Considering the imputation is based on the clustering results, *RC* can indirectly show the correctness of clustering algorithm. The larger *RC* is, the more accurate the algorithm is.

(2) Root Mean Standard Squared error *RMSSE*

For the numeric attributes, we suppose missing original values of numeric attributes are $V_1, V_2, \cdots, V_{\gamma*q*n}$ and imputed values based on clustering algorithms are $V_1', V_2', \cdots, V_{\gamma*q*n}'$. The root mean standard squared error *RMSSE* can be described as follows:

$$RMSSE = \sqrt{\frac{\sum_{i=1}^{\gamma*q*n}(\Delta_i)^2}{\gamma*q*n}} \tag{10}$$

$$\Delta_i = \begin{cases} \frac{V_i - V_i'}{Max - Min} & V_i' \neq "*" \wedge Max \neq Min \\ 1 & V_i' = "*" \\ 0 & V_i' \neq "*" \wedge Max = Min \end{cases} \quad (11)$$

Where, *Max* and *Min* correspondingly mean the maximum and minimum values of that attribute with missing numeric value(s).

(3) Clustering Accuracy *r*

For the clustering results, quality is measured by the clustering accuracy *r* defined as:

$$r = \frac{\sum_{i=1}^{k} c_i}{n} \quad (12)$$

Where c_i is the number of objects occurring in their correct cluster and *k* is the number of clusters.

3.2. Experimental Results

During the experiment, the described algorithms are written and carried out in C++. In order to evaluate the accuracy of improved k-prototypes algorithm, we carry out 100 runs of traditional k-prototypes, k-prototypes+*ISMD* and improved k-prototypes respectively with each missing rate of experimental data sets.

According to formula (9), the *RC* results in the Thyroid disease data set are shown in Figure 1. It is notable that the accurate rate in categorical attributes *RC* of the improved k-prototypes algorithm is higher than initial one.

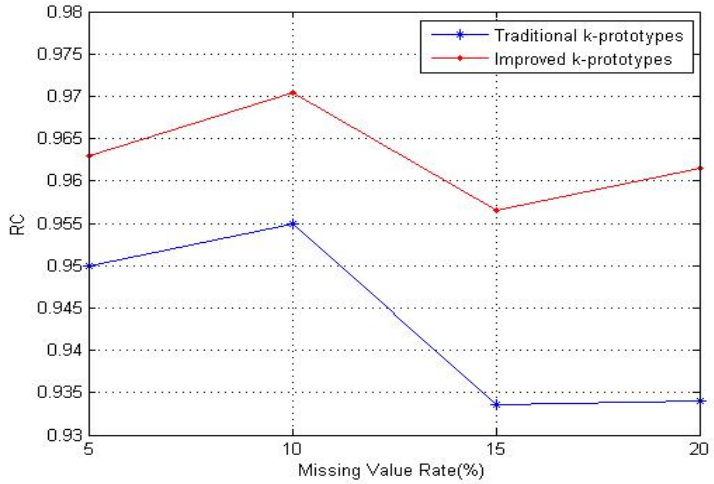

Figure 1. Comparison between Improved k-prototypes and Traditional one in *RC*

As shown in Figure 1, the accurate rate in categorical attributes *RC* of improved k-prototypes is higher than the traditional one with the increase of missing data. The mean *RC* of the former traditional k-prototypes is 0.9431, while that of the improved algorithm is 0.9628. The maximum distance around 20% between them in *RC* is 0.0274 and the improved one has improved by 2.09%, which performs better than the former one.

According to formula (10), the Figure 2 displays the *RMSSE* results in the Thyroid disease data set, which shows that the improved algorithm is a bit better than initial one.

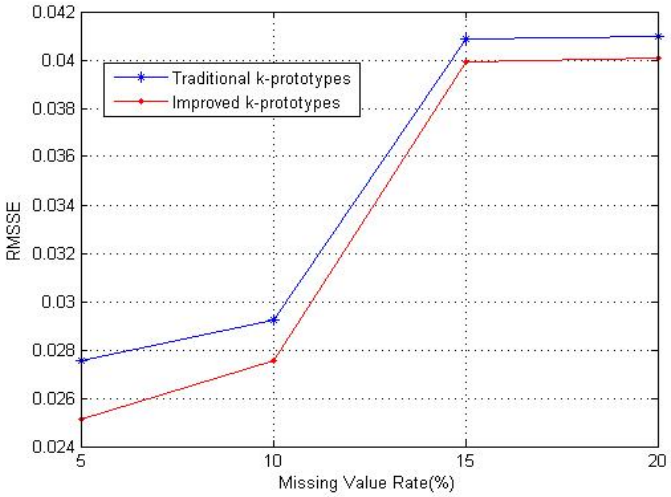

Figure 2. Comparison between Improved k-prototypes and Traditional One in *RMSSE*

From another perspective, the clustering accuracy *r* of traditional k-prototypes, k-prototypes+*ISMD* and improved k-prototypes on the Thyroid disease data is presented in Figure 3 depending on formula (12).

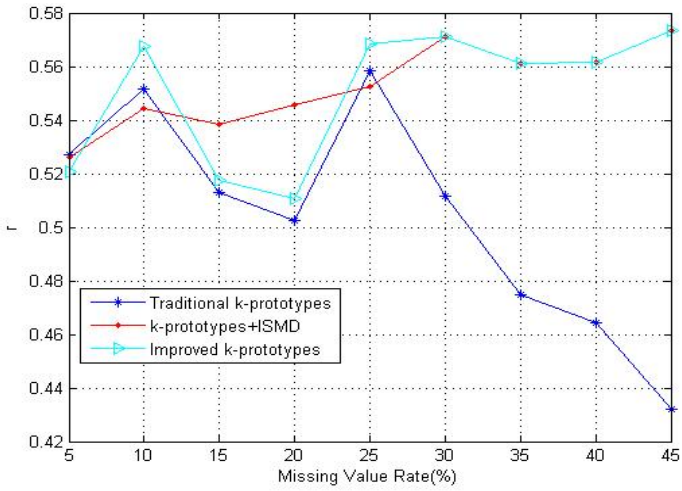

Figure 3. Comparison among Traditional k-prototypes, k-prototypes+*ISMD*, and Improved One in *r*

As shown in Figure 3, with the increase of missing rate, the improved k-prototypes algorithm performs a bit better than k-prototypes+*ISMD* algorithm, obviously better than the traditional one when the missing value rate is above 25%. That's to say, for high missing rate ratio, the improved k-prototypes significantly improves the effectiveness of clustering than traditional one, but not obviously better than k-prototypes+*ISMD*.

To sum up, the improved k-prototypes algorithm has advantage over other algorithms in *RC*, *RMSSE* and *r*. The above experimental results demonstrate that the proposed algorithm not only clusters more accurately on incomplete data but also is effective in filling missing data. Due to the computation of incomplete objects' mixed dissimilarity degree, the improved k-prototypes algorithm doesn't require filling incomplete data in advance before clustering. Meanwhile, the proposed k-prototypes algorithm has improved the calculating method of selecting initial *k* prototypes based on the nearest neighbors.

4. Conclusions and Discussion

This paper proposes the improved k-prototypes algorithm for incomplete data sets with mixed numeric and categorical attributes, which needs us to make comprehensive use of initial cluster prototypes and dissimilarity measure computation, and gains a relatively reasonable clustering and imputation results. Our several comparative tests verify the imputation validity and clustering accuracy of improved k-prototypes algorithm. The main conclusions are as follows:

1) Present an improved formula *(ISMD)* for computing mixed dissimilarity degree of incomplete objects, which not only facilitates to deal with mixed numeric and categorical attributes, but also extends to handle missing data.

2) Cluster first, then impute. The new dissimilarity measure computation takes into account missing data, with no need to impute missing data with means or modes before clustering, which decreases an estimation that may cause error.

3) In order to overcome the shortcomings of random selection of initial k-prototypes in traditional k-prototypes algorithm, this paper chooses *k* initial prototypes based on the idea of the nearest neighbors, improving the clustering accuracy.

4) Impute the missing data based on the clustering result, which contributes to missing data imputation problem to some extent.

5) Compared with traditional k-prototypes and k-prototypes+*ISMD*, the improved k-prototypes algorithm shows higher clustering accuracy in terms of *RC*, *RMSSE* or *r*.

Acknowledgements

This research was partly supported by the National Natural Science Foundation of China under Grant No.71271027 and by the Fundamental Research Funds for the Central Universities of China under Grant No. FRF-TP-10-006B and by the Research Fund for the Doctoral Program of Higher Education under Grant No. 20120006110037.

References

[1] H. Zhexue, "Extension to the K-means algorithm for clustering large data sets with categorical values", Data Mining and Knowledge Discovery, **(1998)**, pp. 283-304.
[2] T. Covões and E. Hruschka, "A study of K-Means-based algorithms for constrained clustering", Intelligent Data Analysis, vol. 17, no. 3, **(2013)**, pp. 485-505.
[3] H. Zhexue, "Clustering large data sets with mixed numeric and categorical values", Proceedings of the 1th pacific-Asia Conference on Knowledge Discovery & Data Mining. Singapore: World Scientific, **(1997)**, pp. 21-34.
[4] W. Qian, W. Cheng and F. Zhenyuan, "Summary of k-means clustering algorithm", Electronic Design Engineering, vol. 20, no. 7, **(2012)**, pp. 21-24.

[5] Z. Lijiang and H. Yongqing, "Improved Clustering Algorithm for Mixed Numeric and Categorical Values", Journal of Guangxi Normal University (Natural Science Edition), **(2007)** April.
[6] W. Mengshu and W. Xizhi, "An improved algorithm for fuzzy k-prototypes algorithm", Statistics and Decision, **(2008)** May.
[7] C. Dan and W. Zhenhua, "A K-prototypes Algorithm Based on Improved Initial Center Points", Computer Knowledge and Technology, **(2010)** November.
[8] Z. Tao and L. Huiling, "Clustering algorithm in data mining research progress", Computer Engineering and Applications, vol. 48, no. 12, **(2012)**.
[9] F. Comellas and A. Miralles, "A fast and efficient algorithm to identify clusters in networks. Applied Mathematics and Computation, vol. 217, no. 5, **(2010)**, pp. 2007-2014.
[10] M. Jinhui, "Clustering algorithm for high-dimensional mixed data sets", Inner Mongolia University of Science and Technology, **(2011)**.
[11] F. Xiaodong, W. Sen and L. Yanchi, "Imputing Missing Value for Mixed Numeric and Categorical Attributes Based on Incomplete Data Hierarchical clustering", KSEM, **(2011)**, pp. 414-424.
[12] B. Tian, J. Jinzhao and H. Jialiang, "New clustering method of mixed-attribute data", Journal of Jilin University, vol. 43, no. 1, **(2013)**.
[13] C. Sotirios, "A fuzzy c-means-type algorithm for clustering of deal with mixed numeric and categorical attributes employing a probabilistic dissimilarity functional", Expert Systems with Applications, vol. 38, no. 7, **(2011)**, pp. 8684-8689.
[14] W. Lei and L. Tianrui, "Matrix-Based Computational Method for Upper and Lower Approximations of Rough Sets", Pattern Recognition and Artificial Intelligence, vol. 24, no. 6, **(2011)**.
[15] M. Weimin and S. Bingzhen, "Probabilistic rough set over two universes and rough entropy", International Journal of Approximate Reasoning, vol. 53, no. 4, **(2012)**, pp. 608-619.
[16] W. Fengmei and H. Lixia, "A Missing Data Imputation Method Based on Neighbor Rules", Computer Engineering, vol. 38, no. 21, **(2012)**.
[17] X. Fang and Z. Guizhu, "Clustering algorithm based on Modified Shuffled Frog Leaping Algorithm and K-means", Computer Engineering and Applications, vol. 49, no. 1, **(2013)**, pp. 176-180.
[18] J. Twisk, M. de Boer, W. de Vente and M. Heymans, "Multiple imputation of missing values was not necessary before performing a longitudinal mixed-model analysis", Journal of Clinical Epidemiology, vol. 66, no. 9, **(2013)**, pp. 1022-1028.
[19] W. Guoyin, "Expansion in the theory of rough set in incomplete information system", Journal of computer research and development, vol. 33, no. 10, **(2002)**, pp. 1239-1240.

Author

Wu Sen received the Ph.D. degree in control theory and control engineering from the University of Science and Technology Beijing, Beijing, China, in 2002.

She is currently a Professor with the Department of Management Science and Engineering, Dongling School of Economics and Management, University of Science and Technology Beijing. Her general areas of research are data mining, complex networks, and personalized recommendation, with a special interest in solving the problems raised from the emerging data-intensive applications. She is currently the Principal Investigator of a NSFC project and a MOE project. She has published four books and over 70 papers in refereed conference proceedings and journals. She has also been a Reviewer for NSFC proposals, the leading academic journals, and many international conferences in her area.

Prof. Wu is a Board Member of the Asian Association of Management Science and Applications. She was a recipient of "I love my teacher—the most excellent teacher in my heart" Honor at the University of Science and Technology Beijing.

