香港浸會大學
HONG KONG BAPTIST UNIVERSITY

計算機科學系
Department of Computer Science

Report production coordinator: Suan Choi, suan@comp.hkbu.edu.hk, +852 3411 7079
Report Web Site: http://www.comp.hkbu.edu.hk/tech-report

A Unified Metric for Categorical and Numerical Attributes in Data Clustering

Yiu-ming Cheung, Hong Jia

COMP-11-001

Release Date: July 22, 2011

Department of Computer Science, Hong Kong Baptist University

Abstract

Most of the existing clustering approaches concentrate on purely numerical or categorical data only, but not the both. In general, it is a nontrivial task to perform clustering on mixed data composed of numerical and categorical attributes because there exists an awkward gap between the similarity metrics for categorical and numerical data. This paper therefore presents a unified metric for data clustering, in which the attributes are in either one of the three types: numerical, categorical, and their both. We firstly present a general clustering framework based on the concept of object-cluster similarity. Then, a unified metric of object-cluster similarity is presented. Finally, an iterative clustering algorithm is developed, which is directly applicable to the three data types stated above without any adjustment. Experimental results show the efficacy of the proposed approach.

Report Cover Designer: So Kin Ming

A Unified Metric for Categorical and Numerical Attributes in Data Clustering

Yiu-ming Cheung
Department of Computer Science
Hong Kong Baptist University
Hong Kong SAR, China
ymc@comp.hkbu.edu.hk

Hong Jia
Department of Computer Science
Hong Kong Baptist University
Hong Kong SAR, China
hjia@comp.hkbu.edu.hk

Abstract

Most of the existing clustering approaches concentrate on purely numerical or categorical data only, but not the both. In general, it is a nontrivial task to perform clustering on mixed data composed of numerical and categorical attributes because there exists an awkward gap between the similarity metrics for categorical and numerical data. This paper therefore presents a unified metric for data clustering, in which the attributes are in either one of the three types: numerical, categorical, and their both. We firstly present a general clustering framework based on the concept of object-cluster similarity. Then, a unified metric of object-cluster similarity is presented. Finally, an iterative clustering algorithm is developed, which is directly applicable to the three data types stated above without any adjustment. Experimental results show the efficacy of the proposed approach.

1 Introduction

To discover the naturally group structure of objects represented in numerical or categorical attributes [5], clustering analysis has been widely applied to a variety of scientific areas, e.g. computer science [1], informatics [2], biology [3], market management [4], and so on. Traditionally, clustering analysis concentrates on purely numerical data only. The typical clustering algorithms include the k-means [6] and EM algorithm [7]. Since the objective functions of these two algorithms are both numerically defined, they are not essentially applicable to the data sets with categorical attributes. Under the circumstances, a straightforward way to overcome this problem is to transform the categorical values into numerical ones, e.g. the binary strings, and then applies the aforementioned numerical-value based clustering methods. Nevertheless, such a method has ignored the similarity information embedded in the categorical values and cannot faithfully reveal the similarity structure of the data sets [8, 9]. Hence, in the past decades, a number of research works have been done towards categorical or even mixed data that are composed of numerical and categorical attributes.

Roughly, the existing approaches dealing with categorical data sets can be summarized into the four categories. The first category of the methods is based on the perspective of similarity. For example, based on Goodall similarity metric [10] that assigns a greater weight to uncommon feature value matching in similarity computations without assuming the underlying distributions of the feature values, Li and Biswas [11] presented the Similarity Based Agglomerative Clustering (SBAC) algorithm. This method has a good capability of dealing with the mixed numeric and categorical attributes, but its computation is quite laborious. Furthermore, ROCK algorithm proposed by Guha et al. [12] is an agglomerative hierarchical clustering procedure based on the concepts of neighbors and links. The desired cluster structure is obtained by merging the clusters sharing a pre-assigned number of neighbors or links gradually. In general, the performance of this algorithm is sensitive to a parameter whose setting is, however, very difficult. Also, the computation of links between objects is quite time-consuming [13]. Beside the similarity concepts, the second category is based

on graph partitioning. A typical example is the CLICKS algorithm [14]. It encodes a data set into a weighted graph structure, where each weighted vertex stands for an attribute value and two nodes are connected if there is a sample in which the corresponding attribute values co-occur. However, as pointed out in [13], this algorithm is not applicable to the clustering task if data sets are high-dimensional or a number of candidates for each attribute becomes large. Moreover, its performance also depends upon a set of parameters whose tuning is very difficult from the practical viewpoint. The third category is entropy-based methods. For example, the COOLCAT algorithm proposed by Barbara et al. [15] utilizes the information entropy to measure the closeness between objects and presents a scheme to find a clustering structure via minimizing the given entropy criterion. Furthermore, a scalable algorithm called LIMBO [16], which is proposed based on the Information Bottleneck (IB) framework [17], employs the concept of mutual information to find a clustering with minimum information loss. Often, the performance of this method relies on the setting of a user-defined parameter. The last category of approaches attempts to give a distance metric between categorical values so that the distance-based clustering algorithm (e.g. the k-means) can be directly adopted. Along this line, the k-modes algorithm is the most cost-effective one proposed by Huang [25, 19, 20]. In this method, the distance between two categorical values is defined as 0 if they are the same, and 1 otherwise. This algorithm is simple and scalable to a large data set. Nevertheless, it has two potential problems: (1) Its performance is sensitive to the initial modes and the definition of categorical distance has not considered the distribution of different categorical values. Accordingly, some improved variants of k-modes algorithm have been therefore exploited, e.g. see [21, 22, 23, 24]. (2) The k-modes algorithm is only applicable to purely categorical data while its complete form, namely k-prototype algorithm [25], is proposed for mixed data clustering, i.e. data clustering with the both of numerical and categorical data. In the k-prototype, different metrics are adopted for numerical and categorical attributes and a user-defined parameter is utilized to control the proportions of numerical distance and categorical distance. Nevertheless, the various setting of this parameter will lead to a totally different clustering result. Recently, Ahmad and Dey [26] has improved the Huang's categorical distance metric and proposed a new mixed data clustering algorithm, in which the cost function weights each numerical attribute by the different values. However, the computation of the categorical distance in this method is quite time-consuming.

In this paper, we will propose a unified clustering approach for both categorical and numeric data sets. Firstly, we present a general clustering framework based on the concept of object-cluster similarity. Then, a unified metric used in data clustering is proposed. Under this metric, the object-cluster similarity for either categorical or numerical attributes has a uniform criterion. Hence, transformation and parameter adjustment between categorical and numerical values are circumvented. Moreover, analogous to the k-means, an iterative algorithm is introduced to implement the data clustering. Experimental results on different benchmark data sets have shown the effectiveness and efficiency of the proposed method.

The rest of the paper is organized as follows. Section 2 gives a general clustering framework based on object-cluster similarity. Also, we present a unified metric of this similarity for categorical, numerical, and mixed attributes. Section 3 describes an algorithm for clustering implementation within the proposed framework. The computational complexity of this algorithm is analyzed as well. Section 4 empirically investigates the performance of the proposed approach in comparison with the existing methods. Finally, we draw a conclusion in Section 5.

2 Object-cluster Similarity Metric

The general task of clustering is to classify the given objects into several clusters such that the similarities between objects in the same group are high while the similarities between objects in different groups are low [27], clustering a set of n objects into k different clusters, denoted as C_1, C_2, ..., C_k, can be formulated to find the optimal $\mathbf{Q}^*$ via the following objective function:

$$\mathbf{Q}^* = \arg\max_{\mathbf{Q}} F(\mathbf{Q}) = \arg\max_{\mathbf{Q}} [\sum_{j=1}^{k} \sum_{i=1}^{n} q_{ij} s(\mathbf{x}_i, C_j)] \tag{1}$$

where $s(\mathbf{x}_i, C_j)$ is the similarity between object $\mathbf{x}_i$ and Cluster C_j, and $\mathbf{Q} = (q_{ij})$ is an $n \times k$ partition matrix satisfying

$$q_{ij} \in [0,1],\ 0 < \sum_{i=1}^{n} q_{ij} < n,\ \text{and}\ \sum_{j=1}^{k} q_{ij} = 1, \qquad i = 1,2,...,n, j = 1,2,...k. \tag{2}$$

Supposing the clustering is hard decision, i.e. the degree that an object i belongs to Cluster j is either 0 or 1, we have

$$q_{ij} = \begin{cases} 1, \text{ if } \mathbf{x}_i \in C_j \\ 0, \text{ otherwise.} \end{cases} \tag{3}$$

Evidently, the desired clusters can be obtained by Eq. (1) as long as the metric of object-cluster similarity is determined. In the following sub-sections, we shall therefore study the similarity metric.

2.1 Similarity Metric for Categorical Attributes

Suppose the set of objects are represented by a set of categorical attributes $A_1, A_2, ..., A_d$, where d is the dimension of the data. The value domain associated with each attribute A_r, denoted as $dom(A_r)$ ($r = 1,2,...,d$), contains all the possible values can be chosen by this attribute. For categorical attributes, the value domains are finite and unordered, i.e., the domain of A_r has m_r elements with $dom(A_r) = \{a_{r1}, a_{r2}, \ldots, a_{rm_r}\}$ and for any $a, b \in dom(A_r)$, either $a = b$ or $a \neq b$ [24]. Then, object $\mathbf{x}_i$ can be denoted by a vector $(x_{i1}, x_{i2}, ..., x_{id})^T$, where $x_{ir} \in dom(A_r)$, $r = 1,2,...,d$, and T is the transpose operator of a matrix.

Definition 1: The similarity between categorical data $\mathbf{x}_i$ and Cluster C_j, $i \in \{1,2,...,n\}$, $j \in \{1,2,...,k\}$, is defined as:

$$s(\mathbf{x}_i, C_j) = \sum_{r=1}^{d} w_r s(x_{ir}, C_j) \text{ with } \sum_{r=1}^{d} w_r = 1. \tag{4}$$

That is, the object-cluster similarity for categorical data is the weighted summation of the similarity between the cluster and each attribute value. The weight factor associate with each attribute describes the importance of an attribute and is utilized to control the contribution of attribute-cluster similarity to object-cluster similarity.

Definition 2: The similarity between an attribute value x_{ir} and Cluster C_j, $i \in \{1,2,...,n\}$, $r \in \{1,2,...,d\}$, $j \in \{1,2,...,k\}$, is defined as:

$$s(x_{ir}, C_j) = \frac{\sigma_{A_r = x_{ir}}(C_j)}{\sigma_{A_r \neq NULL}(C_j)}, \tag{5}$$

where $\sigma_\beta(C_j)$ counts the number of objects (also called *instances* hereinafter) in Cluster C_j with the constraint β.

From **Definition 2**, we can find that this metric of attribute-cluster similarity has the following properties:

(1) $0 \leq s(x_{ir}, C_j) \leq 1$;

(2) $s(x_{ir}, C_j) = 1$ only if all the instances belonging to Cluster C_j have the value x_{ir} for attribute A_r, and $s(x_{ir}, C_j) = 0$ only if no instance belonging to Cluster C_j has the value x_{ir} for attribute A_r.

According to **Definition 1** and **Definition 2**, the object-cluster similarity for categorical data can be therefore calculated by

$$s(\mathbf{x}_i, C_j) = \sum_{r=1}^{d} w_r s(x_{ir}, C_j) = \sum_{r=1}^{d} w_r \frac{\sigma_{A_r = x_{ir}}(C_j)}{\sigma_{A_r \neq NULL}(C_j)}, \tag{6}$$

where $s(\mathbf{x}_i, C_j) \in [0,1]$, $i \in \{1,2,\ldots,n\}$, and $j \in \{1,2,\ldots,k\}$.

2.2 Calculation of Attribute Weights

From the view point of information theory, the significance of an attribute can be regarded as the inhomogeneity degree of the data set with respect to this attribute. Furthermore, it is described in [28] that if the information content of an attribute is high, then the inhomogeneity of the data set is also high for this attribute. Hence, the importance of an attribute A can be quantified by the following entropy metric:

$$H_A = -\int p(x(A)) \log(p(x(A)))dx(A), \tag{7}$$

where $x(A)$ is the value of attribute A, and $p(x(A))$ is the distribution function of the data along this dimension. For categorical attributes, since the values are discrete and independent, we can count the frequency of each attribute value to estimate its probability. Consequently, the importance of any categorical attribute A_r ($r \in \{1, 2, ..., d\}$) can be calculated by

$$H_{A_r} = -\sum_{t=1}^{m_r} p(a_{rt}) \log p(a_{rt}) \text{ with } p(a_{rt}) = \frac{\sigma_{A_r=a_{rt}}(X)}{\sigma_{A_r \neq NULL}(X)}, \tag{8}$$

where $a_{rt} \in dom(A_r)$ and X is the whole data set. Furthermore, according to Eq. (8), the more different values an attribute has, the higher its significance is. However, in practice, an attribute with too many different values may have little contribution to clustering. For example, the ID number of an instance is useless for clustering analysis. Hence, Eq. (8) can be further modified by

$$H_{A_r} = -\frac{1}{m_r} \sum_{t=1}^{m_r} p(a_{rt}) \log p(a_{rt}). \tag{9}$$

That is, the importance of an attribute is quantified by its average entropy over each attribute value. The weight of each attribute is then computed as

$$w_r = \frac{H_{A_r}}{\sum_{t=1}^{d} H_{A_t}}, r = 1, 2, ..., d. \tag{10}$$

Specially, if we assume that all the attributes have the same contribution to the clustering structure of data, the weight of each attribute will become a constant, i.e. $w_r = 1/d$ with $r = 1, 2, \ldots, d$. Subsequently, the object-cluster similarity in Eq. (4) will be simplified to

$$s(\mathbf{x}_i, C_j) = \frac{1}{d} \sum_{r=1}^{d} \frac{\sigma_{A_r=x_{ir}}(C_j)}{\sigma_{A_r \neq NULL}(C_j)}. \tag{11}$$

2.3 Similarity Metric for Numerical Attributes

If the objects are represented by numerical attributes only, we can denote $\mathbf{x}_i$ as $(x_{i1}, x_{i2}, ..., x_{id})$ with $x_{ir} (r \in \{1, 2, ..., d\})$ belongs to $\mathbf{R}$. Under this situation, the distance between each object can be numerically calculated.

Definition 3: The similarity between numeric data $\mathbf{x}_i$ and Cluster C_j, $i \in \{1, 2, ..., n\}$, $j \in \{1, 2, ..., k\}$, is given by

$$s(\mathbf{x}_i, C_j) = \frac{\exp(-0.5 \left\| \mathbf{x}_i - c_j \right\|^2)}{\sum_{t=1}^{k} \exp(-0.5 \left\| \mathbf{x}_i - c_t \right\|^2)}, \tag{12}$$

where c_j is the center of Cluster C_j.

The values of this object-cluster similarity also fall into the interval [0, 1]. Actually, it can be derived that this similarity metric is equivalent to the posterior probability of $\mathbf{x}_i$ belonging to Cluster C_j provided that the probability density function of each object is a mixture of standard normal distribution with the equal mixture coefficients.

2.4 Similarity Metric for Mixed Attributes

This sub-section will study the similarity metric for data mixed with categorical and numerical attributes. Suppose the mixed data $\mathbf{x}_i$ consists of d_c categorical attributes and d_u numerical attributes ($d_c + d_u = d$). $\mathbf{x}_i$ can be therefore denoted as $[\mathbf{x}_i^c, \mathbf{x}_i^u]$ with $\mathbf{x}_i^c = (x_{i1}^c, x_{i2}^c, ..., x_{id_c}^c)$ and $\mathbf{x}_i^u = (x_{i1}^u, x_{i2}^u, ..., x_{id_u}^u)$, where $\mathbf{x}_i^c$ and $\mathbf{x}_i^u$ stand for the categorical part and numerical part of $\mathbf{x}_i$, respectively. Based on the previous definitions, the object-cluster similarity between mixed data $\mathbf{x}_i$ and Cluster C_j can be defined as

$$s(\mathbf{x}_i, C_j) = \sum_{r=1}^{d_c} w_r \frac{\sigma_{A_r=x_{ir}^c}(C_j)}{\sigma_{A_r \neq NULL}(C_j)} + w_{d_c+1} \frac{\exp(-0.5 \left\| \mathbf{x}_i^u - c_j^u \right\|^2)}{\sum_{t=1}^{k} \exp(-0.5 \left\| \mathbf{x}_i^u - c_t^u \right\|^2)} \tag{13}$$

with $\sum_{r=1}^{d_c+1} w_r = 1$, where c_j^u denotes the center of numerical attributes in Cluster C_j. In Eq. (13), it can be seen that the numerical attributes are included as a whole in the Euclidean distance metric, we treat them as indivisible components and assign one weight only to it. Hence, we have $d_c + 1$ attribute weights in total, whose summation should be equal to 1. Under the circumstances, we set the attribute weights at:

$$w_{d_c+1} = \frac{1}{d_c + 1}, \text{ and } w_r = \frac{d_c H_{A_r}}{(d_c + 1) \sum_{t=1}^{d} H_{A_t}}, r = 1, 2, ..., d_c. \tag{14}$$

That is, the total weights of numerical part and categorical part are $\frac{1}{d_c+1}$ and $\frac{d_c}{d_c+1}$, respectively. Moreover, since the actual weight of each categorical attribute is adjusted by its importance, analogous to Eq. (10), we can get the weights expressed by Eq. (14) for mixed attributes.

Before closing this section, please note that the defined similarities for categorical and numerical attributes in Eq. (13) are in the same scale. Hence, unlike k-prototype method, additional parameters to control the proportions of numerical and categorical distances are not needed any more.

3 Clustering Algorithm

This paper concentrates on hard partition only, i.e., $q_{ij} \in \{0, 1\}$, although it can be easily extended to the soft partition. Hence, given a set of n objects, the optimal $\mathbf{Q}^* = \{q_{ij}^*\}$ in Eq. (1) can be given by

$$q_{ij}^* = \begin{cases} 1, \text{ if } s(\mathbf{x}_i, C_j) \geq s(\mathbf{x}_i, C_r), 1 \leq r \leq k, \\ 0, \text{ otherwise.} \end{cases} \tag{15}$$

According to Eq. (15), similar with the k-means, an iterative algorithm can be conducted as follows to implement the clustering analysis:

Step 1 : Given n d-dimensional objects denoted as $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_n$, calculate the importance of each categorical attribute according to Eq. (8), if applicable.

Step 2 : Randomly select k initial objects, one for each cluster.

Step 3 : Choose an object $\mathbf{x}_i$, assign it to Cluster C_j such that $s(\mathbf{x}_i, C_j)$ is maximized, where $s(\mathbf{x}_i, C_j)$ in Eq. (6), Eq. (12) and Eq. (13) is calculated for categorical, numerical, and mixed data, respectively.

Step 4 : Update the cluster information, such as the frequency of each categorical value and the center of numerical part for each cluster.

Step 5 : Repeat **Step 3** and **Step 4** until no objects has changed cluster membership.

Additionally, in order to conveniently update the cluster information in **Step 4**, two auxiliary matrices for each cluster are maintained. One matrix is to record the frequency of each categorical value occurring in this cluster, and the other matrix stores the mean vector of the numerical parts of all objects belonging to this cluster.

We further give the time complexity analysis of the proposed algorithm. The computation cost of **Step 1** is $O(mnkd_c)$. For each iteration, the cost of **Step 2 - Step 5** is $O(mnkd_c + nkd_u)$, where m is the average number of different values can be chosen by each categorical attribute. Hence, the total time cost of this algorithm is $O(t(mnkd_c + nkd_u))$, where t is the number of iterations. From the practical viewpoint, we often have $k << n$, $m << n$ and $t << n$. Subsequently, the time complexity of this algorithm is $O(n)$. Hence, the proposed algorithm is efficient for data clustering, particularly for a large data set.

4 Experiments

To investigate the effectiveness of the proposed approach for data clustering, we applied it to various categorical and mixed data sets obtained from UCI Machine Learning Data Repository[1] and compared its performance with some other popular methods. The algorithm was coded with MATLAB and all the experiments were implemented by a desktop PC computer with Intel(R) Core(TM)2 Quad CPU, 2.40 GHz main frequency, and 4GB DDR2 667 RAM.

4.1 Experiments on Categorical Data Sets

The information of utilized data sets is as follows:

- *Small Soybean Database*: There are 47 instances characterized by 35 multi-valued categorical attributes. According to the different kind of diseases, all the instances should be divided into four groups.
- *Wisconsin Breast Cancer Database*: This data set has 699 instances described by 9 categorical attributes with the values from 1 to 10. Each instance belongs to one of the two clusters labeled by *benign* (contains 458 instances) and *malignant* (contains 241 instances).
- *Congressional Voting Records Data Set*: There are 435 votes based on 16 key features and each vote comes from one of the two different party affiliations: *democrat* (267 votes) and *republican* (168 votes).
- *Zoo Data Set*: This data set consists of 101 instances represented by 16 attributes, in which each instance belongs to one of the 7 animal categories.

In this experiment, the performance of the proposed algorithm was compared with the other three existing approaches for categorical data clustering: k-modes with random initialization (Method I) [25], k-modes with Ng's dissimilarity metric (Method II) [24], and k-modes with initialization using evidence accumulation (Method III) [21]. The clustering accuracy for measuring the clustering results can be estimated by

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

where k is the number of clusters, n stands for the number of instances in the data set, and ai denotes the number of objects that has been correctly assigned to the *i*-th cluster. Consequently, the clustering error is computed as $e = 1 - AC$.

Each algorithm was implemented 100 times on the data sets and the clustering results are summarized in Table 1 and Table 2. Table 1 lists the average error and standard deviation in error obtained by the four different algorithms. From this table, it can be seen that the proposed algorithm has competitive advantage in terms of clustering accuracy and robustness to the initialization compared with the other three methods. Moreover, although Method III is conducted with optimized initial modes, the proposed method with random initialization still gets much smaller standard deviation on the second and third data sets.

Additionally, we further evaluated the convergence speed of the proposed method. Table 2 lists the average epoch number and convergence time over 100 runs cost by the proposed algorithm. Since the k-modes algorithm with random initialization is the fastest one among the other three methods,

[1] http://archive.ics.uci.edu/ml/

Table 1: Comparison of the clustering results of the four different methods on categorical data sets

Data Set	Method I		Method II		Method III		Proposed Method	
	Avg. Error	Error Std	Avg. Error	Error Std	Avg. Error	Error Std	Avg. Error	Error Std
Soybean	0.1589	0.1483	0.0746	0.1351	**0.021**	**0.102**	0.1015	0.1243
Breast Cancer	0.1556	0.1754	0.1497	0.1226	0.132	0.044	**0.0934**	**0.0010**
Vote	0.1395	0.0068	0.1326	0.0063	0.132	0.0071	**0.1214**	**0.0009**
Zoo	0.1644	0.1007	**0.1423**	0.1103	0.166	**0.054**	0.1511	0.1007

Table 2: Comparison of the convergence rate between k-modes and the proposed algorithm

Data Set	Method I (k-modes)			Proposed Method		
	Avg. Epoch No.	Std of Epoch No.	Avg. Time	Avg. Epoch No.	Std of Epoch No.	Avg. Time
Soybean	2.9500	1.0577	0.0161s	1.8300	0.8172	**0.0039**s
Breast Cancer	2.5700	0.6553	0.1048s	2.2100	0.4094	**0.0454**s
Vote	2.4100	0.5336	0.0715s	2.1200	0.3015	**0.0305**s
Zoo	2.8200	0.7834	0.0419s	2.2600	0.8833	**0.0082**s

we only selected it for comparison. It can be seen that the convergence speed of the proposed algorithm is much faster than the k-modes with the improvement of 67% on average in all cases we have tried so far.

4.2 Experiment on Mixed Data Sets

In this experiment, we further investigated the performance of the proposed algorithm on mixed data sets. The selected data sets are as follows:

- *Heart Disease Database*: There are 303 instances characterized by 7 categorical attributes and 6 numeric attributes. All the instances can be grouped into two classes: *healthy* (164 instances) and *sick* (139 instances).
- *Credit Approval Data Set*: This data set has 653 instances described by 9 categorical attributes and 6 numeric attributes. Each instance belongs to one of the two clusters labeled by *positive* (contains 296 instances) and *negative* (contains 357 instances).
- *German Credit Data Set*: This data set contains 1000 instances with 13 categorical attributes and 7 numeric attributes. Each instance is labeled as *good* (700 instances) or *bad* (300 instances).

The performance of the proposed algorithm has been summarized in Table 3. It can be seen that the proposed algorithm is effective for mixed data as well with the satisfactory convergence speed.

5 Conclusions

In this paper, we have proposed a general clustering framework based on object-cluster similarity, through which a unified similarity metric for both categorical and numerical attributes has been presented. Under this new metric, the object-cluster similarity for categorical and numeric data are with the same scale, which is beneficial to clustering analysis on various data types. Moreover, an iterative algorithm has been presented for clustering implementation under the proposed framework. As there is no user-assigned parameter in the proposed algorithm, it is more suitable for the real applications from the practical viewpoint. Experiments have shown the efficacy of this algorithm.

References

[1] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8): 888-905, 2000.

[2] S. Bhatia and J. Deogun. Conceptual clustering in information retrieval. *IEEE Transactions on Systems, Man, and Cybernet, part B: Cybernet*, 28 (3): 427-436, 1998.

Table 3: Clustering result of the proposed algorithm on mixed data sets

Data Set	Proposed Method				
	Avg. Error	Error Std	Avg. Epoch No.	Std of Epoch No.	Avg. Time
Heart Disease	0.1679	0.0032	2.3200	0.4899	0.0036
Credit Approval	0.1858	0.1332	3.5500	1.0188	0.1217
German Credit	0.2915	0.0446	7.5700	3.5083	0.3408

[3] W.H. Au, K.C.C. Chan, A.K.C. Wang, and Y. Wang. Attribute clustering for grouping, selection, and classification of gene expression data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2(2): 83-101, 2005.

[4] J. Hu, B.K. Ray, and M. Singh. Statistical methods for automated generation of service engagement staffing plans. *IBM Journal of Research and Development*, 51(3): 281-293, 2007.

[5] R. Michalski, I. Bratko, and M. Kubat. *Machine learning and data mining: methods and applications*. Wiley, New York, 1998.

[6] J.B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, 1: 281-297, 1967.

[7] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, 39(1): 1-38, 1977.

[8] C.C. Hsu and S.H. Wang. An integrated framework for visualized and exploratory pattern discovery in mixed data. *IEEE Transactions on Knowledge and Data Engineering*, 18(2): 161-173, 2005.

[9] C.C. Hsu. Generalizing self-organizing map for categorical data. *IEEE Transactions on Neural Networks*, 17(2): 294-304, 2006.

[10] D.W. Goodall. A new similarity index based on probability. *Biometric*, 22(4): 882-907, 1966.

[11] C. Li and G. Biswas. Unsupervised learning with mixed numeric and nominal data. *IEEE Transactions on Knowledge and Data Engineering*, 14(4): 673-690, 2002.

[12] S. Guha, R. Rastogi, and K. Shim. ROCK: A robust clustering algorithm for categorical attributes. *Information Systems*, 25(5): 345-366, 2001.

[13] E. Cesario, G. Manco, and R. Ortale. Top-down parameter-free clustering of high-dimensional categorical data. *IEEE Transactions on Knowledge and Data Engineering*, 19(12): 1607-1624, 2007.

[14] M. Zaki and M. Peters. CLICK: Mining subspace clusters in categorical data via k-partite maximal cliques. In *Proceedings of the 21st International Conference on Data Engineering*, pages 355-356, 2005.

[15] D. Barbara', J. Couto, and Y. Li. COOLCAT: An entropy-based algorithm for categorical clustering. In *Proceedings of the 11th ACM Conference on Information and Knowledge Management*, pages 582-589, 2002.

[16] P. Andritsos, P. Tsaparas, R. Miller, and K. Sevcik. LIMBO: Scalable clustering of categorical data. In *Proceedings of Ninth International Conference on Extending Database Technology (EDBT-04)*, pages 123-146, 2004.

[17] N. Tishby, F. C. Pereira, and W. Bialek. The Information Bottleneck Method. In *Proceedings of 37th Annual Allerton Conference on Communication, Control and Computing*, pages 368-377, 1999.

[18] Z. Huang. A fast clustering algorithm to cluster very large categorical data sets in data mining. In *Proceedings of the SIGMOD Workshop on Research Issues on Data Mining and Know ledge Discovery*, pages 1-8, 1997.

[19] Z. Huang. Extensions to the k-modes algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 2(3): 283-304, 1998.

[20] Z. Huang and M. Ng. A note on k-modes clustering. *Journal of Classification*, 20(2): 257-261, 2003.

[21] S.S. Khan and S. Kant. Computation of initial modes for k-modes clustering algorithm using evidence accumulation. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI-07)*, pages 2784-2789, 2007.

[22] F. Cao, J. Liang, and L. Bai. A new initialization method for categorical data clustering. *Expert Systems with Applications*, 36(7): 10223-10228, 2009.

[23] Z. He, S. Deng, and X. Xu. Improving k-modes algorithm considering frequencies of attribute values in mode. In *Proceedings of International Conference on Computational Intelligence and Security*, pages 157-162, 2005.

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

[25] Z. Huang. Clustering large data sets with mixed numeric and categorical values. In *Proceedings of the First Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 21-24, 1997.

[26] A. Ahmad and L. Dey. A k-mean clustering algorithm for mixed numeric and categorical data. *Data & Knowledge Engineering*, 63(2): 503-527, 2007.

[27] A.K. Jain. Data clustering: 50 years beyound k-means. *Pattern Recognition Letters*, 31(8): 651-666, 2010.

[28] J. Basak and R. Krishnapuram. Interpretable hierarchical clustering by constructing an unsupervised decision tree. *IEEE Transactions on Knowledge and Data Engineering*, 17(1):121-132, 2005.