# A Unified Metric for Categorical and Numerical Attributes in Data Clustering

Yiu-ming Cheung[1,2] and Hong Jia[1]

[1] Department of Computer Science and Institute of Computational and Theoretical Studies, Hong Kong Baptist University, Hong Kong SAR, China
[2] United International College, Beijing Normal University - Hong Kong Baptist University, Zhuhai, China
{ymc,hjia}@comp.hkbu.edu.hk

**Abstract.** Most of the existing clustering approaches are applicable to purely numerical or categorical data only, but not 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 general clustering framework based on the concept of object-cluster similarity and gives a unified similarity metric which can be simply applied to the data with categorical, numerical, or mixed attributes. Accordingly, an iterative clustering algorithm is developed, whose efficacy is experimentally demonstrated on different benchmark data sets.

## 1 Introduction

To discover the natural group structure of objects represented in numerical or categorical attributes [1], clustering analysis has been widely applied to a variety of scientific areas. Traditionally, clustering analysis mostly concentrates on purely numerical data only. The typical clustering algorithms include the k-means [2] and EM algorithm [3]. 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 apply 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 [4]. Hence, it is desirable to solve this problem by finding a unified similarity metric for categorical and numerical attributes such that the metric gap between numerical and categorical data can be eliminated. Subsequently, a general clustering algorithm which is applicable to various data types can be presented based on this unified metric.

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

J. Pei et al. (Eds.): PAKDD 2013, Part II, LNAI 7819, pp. 135–146, 2013.

on the concept of object-cluster similarity. Then, a new metric for both of numerical and categorical attributes 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 in data clustering are circumvented. Subsequently, analogous to the framework of k-means, an iterative algorithm is introduced to implement the data clustering. This algorithm conducts an efficient clustering analysis without manually adjusting parameters and is applicable to the three types of data: numerical, categorical, or mixed data, i.e. the data with both of numerical and categorical attributes. Empirical studies have shown the promising results.

## 2 Related Works

Roughly, the existing clustering approaches dealing with data sets which contain categorical attributes can be summarized into the four categories [5]. The first category of the methods is based on the perspective of similarity. For example, based on Goodall similarity metric [6] that assigns a greater weight to uncommon feature value matching in similarity computations without assuming the underlying distributions of the feature values, paper [7] presents 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. Beside the similarity concepts, the second category is based on graph partitioning. A typical example is the CLICKS algorithm [8], which mines subspace clusters for categorical data sets. This novel method 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. It is experimentally demonstrated that CLICKS outperforms ROCK algorithm [9] and scales better for high-dimensional data sets. However, this algorithm is not applicable to data mixed with categorical and numerical attributes and its performance also depends upon a set of parameters whose tuning is quite difficult from the practical viewpoint. The third category is entropy-based methods. For example, the COOLCAT algorithm [10] utilizes the information entropy to measure the closeness between objects and presents a scheme to find a clustering structure via minimizing the expected entropy of clusters. The performance of this algorithm is stable for different data sizes and parameter settings. Nevertheless, this method can only be applied to purely categorical data and cannot handle numerical attributes. The last category of approaches attempts to give a distance metric between categorical values so that the distance-based clustering algorithms (e.g. the k-means) can be directly adopted. Along this line, the most cost-effective one may be the k-prototype algorithm proposed by Huang [11]. In this method, the distance between two categorical values is defined as 0 if they are the same, and 1 otherwise while the distance between numerical values is quantified with Euclidean distance. Subsequently, the k-means paradigm is utilized for clustering. However, since different metrics are adopted for numerical

and categorical attributes, a user-defined parameter is utilized to control the proportions of numerical distance and categorical distance. Nevertheless, various settings of this parameter will lead to a totally different clustering result. A simplified version of k-prototype algorithm namely k-modes [12, 13], which is applicable for purely categorical data clustering, has also been widely utilized due to its satisfactory efficiency, and different improvement strategies have been explored on this method [14–16].

## 3 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 [17]. Therefore, clustering a set of $N$ objects, $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, 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

$$\sum_{j=1}^{k} q_{ij} = 1, \text{ and } 0 < \sum_{i=1}^{N} q_{ij} < N, \tag{2}$$

with

$$q_{ij} \in [0, 1], \ i = 1, 2, \ldots, N, j = 1, 2, \ldots, k. \tag{3}$$

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

### 3.1 Similarity Metric for Mixed Data

This sub-section will study the object-cluster similarity metric for mixed data. Suppose the mixed data $\mathbf{x}_i$ with $d$ different attributes consists of $d_c$ categorical attributes and $d_u$ numerical attributes, i.e. $d_c + d_u = d$. $\mathbf{x}_i$ can be therefore denoted as $[\mathbf{x}_i^{cT}, \mathbf{x}_i^{uT}]^T$ with $\mathbf{x}_i^c = (x_{i1}^c, x_{i2}^c, \ldots, x_{id_c}^c)^T$ and $\mathbf{x}_i^u = (x_{i1}^u, x_{i2}^u, \ldots, x_{id_u}^u)^T$. Then, we have $x_{ir}^u$ $(r = 1, 2, \ldots, d_u)$ belonging to $\mathbf{R}$ and $x_{ir}^c$ $(r = 1, 2, \ldots, d_c)$ belonging to $dom(A_r)$, where $\{A_1, A_2, \ldots, A_{d_c}\}$ are the $d_c$ categorical attributes and $dom(A_r)$ contains all possible values that can be chosen by attribute $A_r$. For categorical attributes, the value domains are finite and unordered, $dom(A_r)$ with $m_r$ elements can be therefore represented with $dom(A_r) = \{a_{r1}, a_{r2}, \ldots, a_{rm_r}\}$.

Firstly, we focus on the difference between categorical attributes and numerical attributes. For categorical attributes, each attribute can usually represent an

important feature of the given object. Therefore, when we conduct classification or clustering analysis, we often investigate the categorical attributes one by one such as Decision Tree method. By contrast, the numerical attributes are often treated as a vector and handled together in clustering analysis. Based on these observations, for the mixed data $\mathbf{x}_i$, the $d_u$ numerical attributes can be treated as a whole but the $d_c$ categorical attributes should be investigated individually. Let the object-cluster similarity between $\mathbf{x}_i$ and cluster $C_j$, denoted as $s(\mathbf{x}_i, C_j)$, be the average of the similarity calculated based on each attribute, we will have

$$\begin{aligned} s(\mathbf{x}_i, C_j) =& \frac{1}{d} s(x_{i1}^c, C_j) + \frac{1}{d} s(x_{i2}^c, C_j) + ... + \frac{1}{d} s(x_{id_c}^c, C_j) + \frac{d_u}{d} s(\mathbf{x}_i^u, C_j) \\ =& \frac{1}{d} \sum_{r=1}^{d_c} s(x_{ir}^c, C_j) + \frac{d_u}{d} s(\mathbf{x}_i^u, C_j). \end{aligned} \tag{4}$$

That is, the similarity between each numerical attribute and the cluster $C_j$ is replaced with the similarity between the cluster and the whole numerical vector $\mathbf{x}_i^u$. Moreover, if we denote the similarity between $\mathbf{x}_i^c$ and $C_j$ as $s(\mathbf{x}_i^c, C_j)$, we can get

$$s(\mathbf{x}_i^c, C_j) = \frac{1}{d_c} \sum_{r=1}^{d_c} s(x_{ir}^c, C_j) = \sum_{r=1}^{d_c} \frac{1}{d_c} s(x_{ir}^c, C_j). \tag{5}$$

Then, (4) can be further rewritten as

$$s(\mathbf{x}_i, C_j) = \frac{d_c}{d} \sum_{r=1}^{d_c} \frac{1}{d_c} s(x_{ir}^c, C_j) + \frac{d_u}{d} s(\mathbf{x}_i^u, C_j) = \frac{d_c}{d} s(\mathbf{x}_i^c, C_j) + \frac{d_u}{d} s(\mathbf{x}_i^u, C_j), \tag{6}$$

where $s(\mathbf{x}_i^c, C_j)$ is actually the similarity on categorical attributes and $s(\mathbf{x}_i^u, C_j)$ is the similarity on numerical attributes. Subsequently, the object-cluster similarity metric can be obtained based on the definitions of $s(\mathbf{x}_i^c, C_j)$ and $s(\mathbf{x}_i^u, C_j)$.

**Similarity Metric for Categorical Attributes.** In (5), we have assumed that each categorical attribute has the same contribution to the calculation of similarity on categorical part. However, from the practical viewpoint, due to the different distributions of attribute values, categorical attributes each often have unequal importance for clustering analysis. In light of this characteristic, (5) should be further modified with

$$s(\mathbf{x}_i^c, C_j) = \sum_{r=1}^{d_c} w_r s(x_{ir}^c, C_j), \tag{7}$$

where $w_r$ is the weight of categorical attribute $A_r$ satisfying $0 \leq w_r \leq 1$ and $\sum_{r=1}^{d_c} w_r = 1$. That is, the object-cluster similarity for categorical part is the weighted summation of the similarity between the cluster and each attribute value. Weight factor $w_r$ describes the importance of each categorical attribute and is utilized to control the contribution of attribute-cluster similarity to object-cluster similarity.

**Definition 1.** *The similarity between a categorical attribute value $x_{ir}^c$ and cluster $C_j$, $i \in \{1, 2, \ldots, N\}$, $r \in \{1, 2, \ldots, d_c\}$, $j \in \{1, 2, \ldots, k\}$, is defined as:*

$$s(x_{ir}^c, C_j) = \frac{\sigma_{A_r = x_{ir}^c}(C_j)}{\sigma_{A_r \neq NULL}(C_j)}, \tag{8}$$

*where NULL refers to empty, and $\sigma_{A_r = x_{ir}^c}(C_j)$ counts the number of objects (also called* instances *hereinafter) that have the value $x_{ir}^c$ for attribute $A_r$ in cluster $C_j$.*

From *Definition 1*, we can find that this metric of attribute-cluster similarity has the following properties:

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

According to (7) and (8), the object-cluster similarity for categorical part can be therefore calculated by

$$s(\mathbf{x}_i^c, C_j) = \sum_{r=1}^{d_c} w_r s(x_{ir}^c, C_j) = \sum_{r=1}^{d_c} w_r \frac{\sigma_{A_r = x_{ir}^c}(C_j)}{\sigma_{A_r \neq NULL}(C_j)}, \tag{9}$$

where $i \in \{1, 2, \ldots, N\}$, and $j \in \{1, 2, \ldots, k\}$.

*Remark 1.* Since $0 \leq s(x_{ir}^c, C_j) \leq 1$ and $\sum_{r=1}^{d_c} w_r = 1$, we have:

$$s(\mathbf{x}_i^c, C_j) = \sum_{r=1}^{d_c} w_r s(x_{ir}^c, C_j) \geq \sum_{r=1}^{d_c} (w_r \cdot 0) = 0,$$

and

$$s(\mathbf{x}_i^c, C_j) = \sum_{r=1}^{d_c} w_r s(x_{ir}^c, C_j) \leq \sum_{r=1}^{d_c} (w_r \cdot 1) = \sum_{r=1}^{d_c} w_r = 1.$$

That is, for any $i \in \{1, 2, \ldots, N\}$ and $j \in \{1, 2, \ldots, k\}$, the value of $s(\mathbf{x}_i^c, C_j)$ will fall into the interval $[0, 1]$.

Next, we discuss how to estimate the importance of each categorical attribute. 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 [18] that if the information content of an attribute is high, the inhomogeneity of the data set is also high for this attribute. Hence, the importance of any categorical attribute $A_r$ ($r \in \{1, 2, \ldots, d_c\}$) can be calculated by

$$H_{A_r} = -\sum_{t=1}^{m_r} p(a_{rt}) \log p(a_{rt}) \tag{10}$$

with

$$p(a_{rt}) = \frac{\sigma_{A_r=a_{rt}}(X)}{\sigma_{A_r \neq NULL}(X)}, \tag{11}$$

where $a_{rt} \in dom(A_r)$, $p(a_{rt})$ is the probability of attribute value $a_{rt}$, $m_r$ is the total number of values that can be chosen by $A_r$ and $X$ is the whole data set. Furthermore, according to (10), 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 instances is unique for each instance, but this information is useless for clustering analysis. Hence, (10) can be further modified with

$$H_{A_r} = -\frac{1}{m_r} \sum_{t=1}^{m_r} p(a_{rt}) \log p(a_{rt}). \tag{12}$$

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

$$w_r = \frac{H_{A_r}}{\sum_{t=1}^{d_c} H_{A_t}}, r = 1, 2, \ldots, d_c. \tag{13}$$

Subsequently, the object-cluster similarity on categorical part can be given by

$$s(\mathbf{x}_i^c, C_j) = \sum_{r=1}^{d_c} \left( \frac{H_{A_r}}{\sum_{t=1}^{d_c} H_{A_t}} \cdot \frac{\sigma_{A_r=x_{ir}^c}(C_j)}{\sigma_{A_r \neq NULL}(C_j)} \right). \tag{14}$$

In practice, for an attribute $A_r$, if all the instances to be classified have the same value $a$, it can be obtained from (12) and (11) that the importance of this attribute will be 0 as $p(a) = 1$ and $log(1) = 0$. Then, the corresponding attribute weight will also be zero and this attribute will have no contribution to the whole clustering learning.

**Similarity Metric for Numerical Attributes.** Since the distance between each vector $\mathbf{x}_i^u$ can be numerically calculated, the similarity metric for numerical attributes can be defined based on the measure of distance. According to [19] and [20], it is a universal law that the distance and perceived similarity between numerical vectors are related via an exponential function as follows:

$$s(\mathbf{x}_A, \mathbf{x}_B) = \exp(-Dis(\mathbf{x}_A, \mathbf{x}_B)), \tag{15}$$

where $Dis$ stands for a distance measure. Moreover, it can be observed that the magnitudes of distances between instances from variant data sets may have a significant difference in practice. To avoid the potential influence of this scenario, we can further use proportional distance instead of absolute distance to estimate the similarity between numerical vectors.

**Definition 2.** *The object-cluster similarity between numerical vector* $\mathbf{x}_i^u$ *and cluster* $C_j$, $i \in \{1, 2, \ldots, N\}$, $j \in \{1, 2, \ldots, k\}$, *is given by*

$$s(\mathbf{x}_i^u, C_j) = \exp\left(-\frac{Dis(\mathbf{x}_i^u, \mathbf{c}_j)}{\sum_{t=1}^{k} Dis(\mathbf{x}_i^u, \mathbf{c}_t)}\right), \tag{16}$$

*where* $\mathbf{c}_j$ *is the center of all numerical vectors in cluster* $C_j$.

It can be seen from *Definition 2* that the values of this similarity metric also fall into the interval [0, 1]. In practice, different distance metrics can be utilized to calculate $Dis(\mathbf{x}_i^u, \mathbf{c}_j)$. For example, if the Minkowski distance is adopted, we shall have:

$$Dis(\mathbf{x}_i^u, \mathbf{c}_j) = \left(\sum_{r=1}^{d_u} |x_{ir}^u - c_{jr}|^p\right)^{1/p}, \tag{17}$$

where $p > 0$ is a constant which characterizes the distance function. A typically special case of (17) is the Euclidean distance with $p = 2$.

Finally, according to (6), (14), and (16), the object-cluster similarity metric for mixed data is defined as

$$s(\mathbf{x}_i, C_j) = \frac{d_c}{d}\sum_{r=1}^{d_c}\left(\frac{H_{A_r}}{\sum_{t=1}^{d_c} H_{A_t}} \cdot \frac{\sigma_{A_r=x_{ir}^c}(C_j)}{\sigma_{A_r \neq NULL}(C_j)}\right) + \frac{d_u}{d}\exp\left(-\frac{Dis(\mathbf{x}_i^u, \mathbf{c}_j)}{\sum_{t=1}^{k} Dis(\mathbf{x}_i^u, \mathbf{c}_t)}\right), \tag{18}$$

where $i = 1, 2, \ldots, N$, $j = 1, 2, \ldots, k$. It can be seen that the defined similarities for categorical and numerical attributes in (18) are in the same scale. Hence, unlike $k$-prototype method, there is no need any more to manually adjust the parameter to control the proportions of numerical and categorical distances for different data sets.

## 4 Iterative 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 in terms of posterior probability. Under the circumstances, given a set of $N$ objects, the optimal $\mathbf{Q}^* = \{q_{ij}^*\}$ in (1) can be given by

$$q_{ij}^* = \begin{cases} 1, & \text{if } s(\mathbf{x}_i, C_j) \geq s(\mathbf{x}_i, C_r), 1 \leq r \leq k, \\ 0, & \text{otherwise.} \end{cases} \tag{19}$$

Therefore, similar to the learning procedure of k-means, an iterative algorithm, denoted as OCIL, can be conducted to implement the clustering analysis as shown in Algorithm 1.

The first step in OCIL algorithm, i.e. Step 1, is a procedure for the calculation of object-cluster similarity. Thus, we can find that the iterative steps of

**Algorithm 1** Iterative clustering learning based on object-cluster similarity metric (OCIL)

**Require:** data set $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, number of clusters $k$
**Ensure:** cluster label $Y = \{y_1, y_2, \ldots, y_N\}$
1: Calculate the importance of each categorical attribute according to (12), if applicable
2: Set $Y = \{0, 0, \ldots, 0\}$ and randomly select $k$ initial objects, one for each cluster
3: **repeat**
4: Initialize $noChange = true$
5: **for** $i = 1$ **to** $N$ **do**
6: $y_i^{(new)} = \arg \max_{j \in \{1,\ldots,k\}} [s(\mathbf{x}_i, C_j)]$
7: **if** $y_i^{(new)} \neq y_i^{(old)}$ **then**
8: $noChange = false$
9: Update the information of clusters $C_{y_i^{(new)}}$ and $C_{y_i^{(old)}}$, including the frequency of each categorical value and the centroid of numerical vectors
10: **end if**
11: **end for**
12: **until** $noChange$ is $true$
13: **return** $Y$

OCIL algorithm is the same as the k-means algorithm and the only difference is the measurement of similarity between object and clusters. Therefore, the effectiveness of the proposed similarity metric can be easily evaluated by comparing OCIL with other similar algorithms, such as k-means and k-prototype. Next, we further give the time complexity analysis of OCIL algorithm. It can be observed that the computation cost of Step 1 is $O(mNd_c)$, where $m$ is the average number of different values that can be chosen by each categorical attribute. For each iteration, the cost of the "**for**" statement is $O(mNkd_c + Nkd_u)$. Hence, the total time cost of this algorithm is $O(t(mNkd_c + Nkd_u))$, where $t$ stands for the number of iterations. From the practical viewpoint, $k$, $m$ and $t$ can be regarded as a constant in most cases. Therefore, the time complexity of this algorithm approaches to $O(dN)$. Hence, the proposed algorithm is efficient for data clustering, particularly for a large data set.

## 5 Experiments

This section is to investigate the effectiveness of the proposed approach to 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 the existing counterparts. Since the proposed method on numerical data degenerates to the k-means algorithm, the effectiveness of OCIL algorithm on numerical data set is transparent. Hence, there is no need to investigate it any more. Each algorithm was coded with MATLAB and all experiments were implemented by

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

a desktop PC computer with Intel(R) Core(TM)2 Quad CPU, 2.40 GHz main frequency, and 4GB DDR2 667 RAM.

Moreover, in our experiments, the clustering accuracy [21] for measuring the clustering performance was estimated by

$$ACC = \frac{\sum_{i=1}^{N} \delta(c_i, map(r_i))}{N},$$

where $N$ is the number of instances in the data set, $c_i$ stands for the provided label, $map(r_i)$ is a mapping function which maps the obtained cluster label $r_i$ to the equivalent label from the data corpus by using the Kuhn-Munkres algorithm, and the delta function $\delta(c_i, map(r_i)) = 1$ only if $c_i = map(r_i)$, otherwise 0. Correspondingly, the clustering error rate is computed as $e = 1 - ACC$.

### 5.1 Performance on Mixed Data Sets

In the following experiments, we will investigate the performance of the proposed algorithm on real data sets in comparison with the existing counterparts. Firstly, experiments were conducted on mixed data and the information of selected data sets is shown in Table 1. The performance of the proposed method has been compared with the k-prototype algorithm [11] and k-means algorithm, whose time complexity are also $O(Nd)$. In k-prototype method, the distance regulation parameter $\gamma$ was set at $0.5\sigma$ [11], where $\sigma$ is the average standard deviation of numerical attributes. When utilizing k-means, the categorical values were transformed into integers in our experiments. Moreover, the Euclidean distance has been adopted as the distance metric of numerical vectors for consistency. Each algorithm has been run 100 times on each data set and the clustering results are summarized in Table 2.

**Table 1.** Statistics of mixed data sets

| Data set | Instance | Attribute ($d_c + d_u$) | Class |
|---|---|---|---|
| Statlog Heart | 270 | 7 + 6 | 2 |
| Heart Disease | 303 | 7 + 6 | 2 |
| Credit Approval | 653 | 9 + 6 | 2 |
| German Credit | 1000 | 13 + 7 | 2 |
| Dermatology | 366 | 33 + 1 | 6 |
| Adult | 30162 | 8 + 6 | 2 |

It can be seen that, both with random initializations, the proposed algorithm OCIL has an obvious superiority in terms of clustering accuracy over the k-prototype and k-means methods. This result shows that, in comparison with numerically representing the distance between categorical values, the proposed similarity metric in this paper is a more reasonable measurement for clustering

**Table 2.** Clustering errors of OCIL on mixed data sets in comparison with k-prototype and k-means

| Data set | K-means | K-prototype | OCIL |
|---|---|---|---|
| Statlog | 0.4047±0.0071 | 0.2306±0.0821 | **0.1716±0.0065** |
| Heart | 0.4224±0.0131 | 0.2280±0.0903 | **0.1644±0.0030** |
| Credit | 0.4487±**0.0016** | 0.2619±0.0976 | **0.2519**±0.0966 |
| German | 0.3290±0.0014 | 0.3289±**0.0006** | **0.3057**±0.0007 |
| Dermatology | 0.7006±**0.0216** | 0.6903±0.0255 | **0.3051**±0.0896 |
| Adult | 0.3869±**0.0067** | 0.3855±0.0143 | **0.3079**±0.0305 |

**Table 3.** Comparison of average convergent time and iterations between k-prototype and OCIL

| Data set | Time | | Iterations | |
|---|---|---|---|---|
| | K-prototype | OCIL | K-prototype | OCIL |
| Statlog | 0.0519s | **0.0516**s | 3.09 | **3.07** |
| Heart | 0.0639s | **0.0576**s | 3.54 | **3.02** |
| Credit | **0.1323**s | 0.1625s | **3.18** | 4.26 |
| German | 0.2999s | **0.2023**s | 5.29 | **3.15** |
| Dermatol | 0.3674s | **0.1888**s | 7.27 | **4.32** |
| Adult | 15.2795s | **9.6774**s | 10.93 | **6.78** |

**Table 4.** Statistics of categorical data sets

| Data set | Instance | Attribute | Class |
|---|---|---|---|
| Soybean | 47 | 35 | 4 |
| Breast | 699 | 9 | 2 |
| Vote | 435 | 16 | 2 |
| Zoo | 101 | 16 | 7 |

**Table 5.** Comparison of clustering errors on categorical data sets

| Data set | H's k-modes | N's k-modes | OCIL |
|---|---|---|---|
| Soybean | 0.1691±0.1521 | **0.0964**±0.1404 | 0.1017±**0.1380** |
| Breast | 0.1655±0.1528 | 0.1356±0.0016 | **0.0934±0.0009** |
| Vote | 0.1387±0.0066 | 0.1345±0.0031 | **0.1213±0.0010** |
| Zoo | 0.2873±0.1083 | 0.2730±**0.0818** | **0.2681**±0.0906 |

analysis on mixed data. Moreover, comparing the average running time of OCIL and k-prototype algorithms listed in Table 3, we can find that, although OCIL needs additional time to calculate the weight of each categorical attribute, its total running time is no more than k-prototype's one. A plausible reason can

be found from Table 3 is that the convergence speed of OCIL is usually faster than k-prototype in most cases we have tried so far. Therefore, the proposed similarity metric is efficient for mixed data clustering.

### 5.2 Performance on Categorical Data Sets

We further investigated the performance of the proposed algorithm on purely categorical data. The information of four different benchmark data sets we utilized has been summarized in Table 4. To conduct comparative studies, we have also implemented the other two existing categorical data clustering algorithms: original k-modes (H's k-modes) [12] and k-modes with Ng's dissimilarity metric (N's k-modes) [16]. In this experiment, each algorithm was conducted with random initializations. Table 5 lists the average value and standard deviation in error obtained by OCIL and the other two algorithms, respectively. It can be seen that the proposed clustering method has competitive advantage in terms of clustering accuracy and robustness compared with the other two methods.

## 6 Conclusion

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 numerical attributes are with the same scale, which is beneficial to clustering analysis on various data types. Subsequently, analogous to k-means method, an iterative algorithm has been introduced to implement the data clustering. The advantages of the proposed method have been experimentally demonstrated in comparison with the existing counterparts.

**Acknowledgment.** The work described in this paper was supported by the Faculty Research Grant of Hong Kong Baptist University with the Project Code: FRG2/11-12/067, and the NSFC under grant 61272366.

## References

1. Michalski, R.S., Bratko, I., Kubat, M.: Machine learning and data mining: methods and applications. Wiley, New York (1998)
2. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297 (1967)
3. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society, Series B (Methodological) 39(1), 1–38 (1977)
4. Hsu, C.C.: Generalizing self-organizing map for categorical data. IEEE Transactions on Neural Networks 17(2), 294–304 (2006)

5. Cesario, E., Manco, G., Ortale, R.: Top-down parameter-free clustering of high-dimensional categorical data. IEEE Transactions on Knowledge and Data Engineering 19(12), 1607–1624 (2007)
6. Goodall, D.W.: A new similarity index based on probability. Biometric 22(4), 882–907 (1966)
7. Li, C., Biswas, G.: Unsupervised learning with mixed numeric and nominal data. IEEE Transactions on Knowledge and Data Engineering 14(4), 673–690 (2002)
8. Zaki, M.J., Peters, M.: Click: Mining subspace clusters in categorical data via k-partite maximal cliques. In: Proceedings of the 21st International Conference on Data Engineering, pp. 355–356 (2005)
9. Guha, S., Rastogi, R., Shim, K.: Rock: A robust clustering algorithm for categorical attributes. Information Systems 25(5), 345–366 (2001)
10. Barbara, D., Couto, J., Li, Y.: Coolcat: An entropy-based algorithm for categorical clustering. In: Proceedings of the 11th ACM Conference on Information and Knowledge Management, pp. 582–589 (2002)
11. Huang, Z.: Clustering large data sets with mixed numeric and categorical values. In: Proceedings of the First Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 21–24 (1997)
12. Huang, Z.: 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 Knowledge Discovery, pp. 1–8 (1997)
13. Huang, Z., Ng, M.K.: A note on k-modes clustering. Journal of Classification 20(2), 257–261 (2003)
14. Khan, S.S., Kant, S.: Computation of initial modes for k-modes clustering algorithm using evidence accumulation. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI 2007), pp. 2784–2789 (2007)
15. He, Z., Deng, S., Xu, X.: Improving k-modes algorithm considering frequencies of attribute values in mode. In: Hao, Y., Liu, J., Wang, Y.-P., Cheung, Y.-m., Yin, H., Jiao, L., Ma, J., Jiao, Y.-C. (eds.) CIS 2005. LNCS (LNAI), vol. 3801, pp. 157–162. Springer, Heidelberg (2005)
16. Ng, M.K., Li, M.J., Huang, J.Z., He, Z.: On the impact of dissimilarity measure in k-modes clustering algorithm. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(3), 503–507 (2007)
17. Jain, A.K.: Data clustering: 50 years beyound k-means. Pattern Recognition Letters 31(8), 651–666 (2010)
18. Basak, J., Krishnapuram, R.: Interpretable hierarchical clustering by constructing an unsupervised decision tree. IEEE Transactions on Knowledge and Data Engineering 17(1), 121–132 (2005)
19. Shepard, R.N.: Toward a universal law of generalization for physical science. Science 237, 1317–1323 (1987)
20. Santini, S., Jain, R.: Similarity measures. IEEE Transactions on Pattern Analysis and Machine Intelligence 21(9), 871–883 (1999)
21. He, X., Cai, D., Niyogi, P.: Laplacian score for feature selection. In: Advances in Neural Information Processing Systems (2005)