algorithms

MDPI

Article

A Similarity Measurement with Entropy-Based Weighting for Clustering Mixed Numerical and Categorical Datasets

Xia Que, Siyuan Jiang, Jiaoyun Yang * and Ning An

School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601, China; quexia@hfut.edu.cn (X.Q.); siyuan@mail.hfut.edu.cn (S.J.); ning.g.an@acm.org (N.A.)
* Correspondence: jiaoyun@hfut.edu.cn

Abstract: Many mixed datasets with both numerical and categorical attributes have been collected in various fields, including medicine, biology, etc. Designing appropriate similarity measurements plays an important role in clustering these datasets. Many traditional measurements treat various attributes equally when measuring the similarity. However, different attributes may contribute differently as the amount of information they contained could vary a lot. In this paper, we propose a similarity measurement with entropy-based weighting for clustering mixed datasets. The numerical data are first transformed into categorical data by an automatic categorization technique. Then, an entropy-based weighting strategy is applied to denote the different importances of various attributes. We incorporate the proposed measurement into an iterative clustering algorithm, and extensive experiments show that this algorithm outperforms OCIL and K-Prototype methods with 2.13% and 4.28% improvements, respectively, in terms of accuracy on six mixed datasets from UCI.

Keywords: automatic categorization; clustering analysis; mixed datasets; similarity measurement; information entropy

check for updates

Citation: Que, X.; Jiang, S.; Yang, J.; An, N. A Similarity Measurement with Entropy-Based Weighting for Clustering Mixed Numerical and Categorical Datasets. *Algorithms* **2021**, *14*, 184. https://doi.org/10.3390/a14060184

Academic Editor: Javier Del Ser Lorente

Received: 22 May 2021
Accepted: 13 June 2021
Published: 15 June 2021

1. Introduction

The main purposes of clustering analyses are to discover the implicit class structure in the data and divide the physical or abstract objects into different classes, where the similarity between a pair of objects in the same class is large and in different classes is small. As a major exploratory data analysis tool, clustering analysis has been widely researched and applied in many fields, such as sociology, biology, medicine, etc. [1–3]. Most current methods are designed to address single dataset types (numerical or categorical). For example, classical clustering methods, such as the k-means algorithm [4,5], the EM algorithm [6], etc., are limited to numerical datasets, while some algorithms are also proposed for clustering categorical datasets [7,8]. However, in the medical and biology fields, many datasets are collected with both numerical and categorical attributes. Hence, many researchers are dedicated to discovering clustering algorithms for mixed types of datasets with categorical and numerical attributes [9,10].

Many unsupervised clustering algorithms for mixed datasets have been proposed over the years, which can be classified into two types. The first type designs different similarity measurements for numerical and categorical data and then calculates the weighted sum of the two parts. For example, the K-Prototypes algorithm [11] for clustering mixed datasets was put forward simply by combining the k-means algorithm and the K-Modes algorithm, which are used for single types of numerical and categorical datasets, respectively. Additionally, the OCIL algorithm proposed by Cheung and Jia [12] is an iterative clustering learning algorithm based on object-cluster similarity metrics.

In the second type, the algorithms transform categorical attributes into numerical ones, and then the algorithms apply clustering methods designed for purely numerical datasets to the transformed dataset or vice versa. The most direct method is to map categorical values into numerical vectors. If a categorical attribute contains n unique values, then each value is mapped into a n-dimensional vector. This strategy increases the dataset dimensions, resulting in higher computational complexity. It could also transform numerical attributes into categorical ones. For instance, SpectralCAT, proposed by David and Averbuch [13], automatically transforms high-dimensional data into categorical data and then applies spectral clustering [14] to reduce the dimensionality of the transformed datasets through automatic non-linear transformations.

When designing clustering algorithms, the similarity or dissimilarity measurement plays an important role. Due to the different nature of numerical attributes and categorical attributes, they should be handled differently. Numerical data use a continuous variable to represent the values of each attribute, and a common distance such as the Euclidean distance usually measures the similarity between numerical objects. However, the values of the categorical data have neither a natural ordering nor a common scale. Due to this distinct nature of these two different data types, methods designed for single-type datasets cannot be applied to other types of datasets. The most direct way is the second of the two types mentioned above. However, this method ignores the similarity information in the categorical attribute values [15]. Therefore, the Hamming distance is used in many dissimilarity measurements. For example, in the K-Prototypes algorithm, the dissimilarity measurement uses the Euclidean distance for the numerical attributes and the Hamming distance for the categorical attributes. This algorithm also controls the contribution of the numerical attributes and the categorical attributes through a user-defined parameter. The K-Prototypes algorithm is simple and easy to implement, so it has been widely used in clustering mixed datasets. However, when implementing similarity measurements for categorical attributes, the Hamming distance is rough, and the clustering result is very sensitive to this parameter in the K-Prototypes algorithm. Subsequently, some improved similarity measurements for categorical attributes are proposed, which are based on the frequency of categorical values, the co-occurrence, and the conditional probability estimate [7,16,17]. Based on these improved similarity measurements for categorical attributes, some combined similarity measurements for both categorical and numerical datasets have been developed. For instance, the OCIL algorithm [12] uses the frequency of categorical object values that occur in the cluster for categorical attributes and the numerical distance for numerical attributes when measuring similarity.

It can be found that each attribute often contributes differently to the desired clustering results in many practical applications, which should be considered when measuring the similarities. For example, we want to cluster a mammographic mass dataset into two groups, corresponding to benign types and malignant types. In this task, the age attribute may play a more important role than the mass density attribute. Therefore, it is very important to identify different attribute contributions to improve the quality of the clustering results. Actually, some researchers have realized this problem and proposed several strategies. However, most research focuses on single-type datasets; e.g., for categorical attributes, the weights could be assigned based on the overall distribution of attribute values [18] or based on the frequency the class center appearances and the average distance between objects and the clustering center [19]. When handling mixed datasets, current algorithms only assign weights for single-type attributes. For example, when the OCIL algorithm [12] measuring the similarity, which only assigns weight for each categorical attribute based on information entropy, and it lets each numerical attribute take the same weight. The result is to weakens the importance of the numerical attributes. Ahmad and Dey proposed an algorithm for mixed datasets by adding weights to only numerical attributes [20]. Actually, both the numerical attributes and the categorical attributes should be evaluated when designing the similarity, and the weight strategy should be applied to both types of attributes in order to simplify the computational complexity.

In this paper, we propose a similarity measurement with entropy-based weighting for mixed datasets with both categorical and numerical attributes. First, a similarity metric for the categorical attributes is designed by assigning a different weight to each attribute based on information entropy theory. Second, we present an automatic categorization technique that transforms numerical data into categorical data, which is achieved by automatically discovering the optimal number of categorizations for each attribute based on the Calinski-Harabasz index. Then, the similarity metric for categorical data can be used to measure the similarity for transformed data. In this way, this similarity measurement can be applied to the mixed dataset containing both numerical and categorical attributes. Subsequently, this similarity measurement with entropy-based weighting is applied to the k-means framework. We accessed several datasets from UCI and compared the proposed algorithm with the OCIL and K-Prototype methods on mixed datasets as well as with the k-means algorithm on numerical datasets. The experimental results show that the iterative clustering algorithm based on the proposed similarity measurement is superior to these three algorithms.

The remainder of this paper is organized as follows. Section 2 introduces the problem formulation and then proposes a similarity measurement with entropy-based weighting for mixed datasets and applies this similarity measurement to the k-means algorithm framework. In Section 3, experiments are conducted to compare the proposed algorithm with three existing methods. Finally, we draw conclusions in Section 4.

2. Methods

2.1. Problem Formulation

Clustering means classifying the given unlabeled objects into several clusters according to certain criteria, so similar objects are classified as one cluster, and dissimilar objects are assigned to different clusters.

For a given mixed dataset X consisting of m objects, denoted as $\{x_1, x_2, \ldots, x_m\}$, suppose X has d_c categorical attributes and d_u numerical attributes. Then, $x_i (1 \leq i \leq m)$ can be denoted as $[x_i^c, x_i^u]$, with $x_i^c = [x_{i,1}^c, x_{i,2}^c, \ldots, x_{i,d_c}^c]$ and $x_i^u = [x_{i,1}^u, x_{i,2}^u, \ldots, x_{i,d_u}^u]$. The requirement is to cluster the dataset X into k different clusters, denoted as $C_1, C_2, \ldots, C_k$, with $C = \{C_1, C_2, \ldots, C_k\}$, and $C_i \cap C_j = \emptyset$, $\cup_{i=1}^{k} C_i = C (i, j = 1, 2, \ldots, k;\ i \neq j)$. The optimal partition matrix T^* can be found through the following objective function:

$$T^* = argmax_T [\sum_{j=1}^{k} \sum_{i=1}^{m} t_{ij} s(x_i, C_j)] \tag{1}$$

where $s(x_i, C_j)$ is the similarity between object x_i and cluster C_j, $T = (t_{ij})$ is an $m \times k$ partition matrix with $t_{ij} \in \{0, 1\}$ and $\sum_{j=1}^{k} t_{ij} = 1$, $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, k$. $t_{ij} = 1$ indicates that object x_i is assigned to cluster j.

According to Equation (1), the clusters can be obtained as long as the metric function of similarity between object x_i and cluster C_j is determined. Because implied information of each attribute is different, the contribution to cluster result is also different., we define a new similarity, in which each attribute is assigned a weight, denoted as w_r, satisfying $0 \leq w_r^c \leq 1, 0 \leq w_r^u \leq 1$ and $\sum_{r=1}^{d_c} w_r^c + \sum_{r=1}^{d_u} w_r^u = 1$. Then the similarity between object x_i and cluster C_j can be measured by the following equation:

$$s(x_i, C_j) = \sum_{r=1}^{d_c} w_r^c s_c(x_{i,r}^c, C_j) + \sum_{r=1}^{d_u} w_r^u s_u(x_{i,r}^u, C_j) \tag{2}$$

where w_r^c and $s_c(x_{i,r}^c, C_j)$ are the weight and similarity on the categorical attribute, respectively, w_r^u and $s_u(x_{i,r}^u, C_j)$ are the weight and similarity on the numerical attribute, respectively. $s_c(x_i^c, C_j) = \sum_{r=1}^{d_c} w_r^c s_c(x_{i,r}^c, C_j)$ represents the similarity on categorical attributes and $s_u(x_i^u, C_j) = \sum_{r=1}^{d_u} w_r^u s_u(x_{i,r}^u, C_j)$ represents the similarity on numerical attributes. In the following sections, we study how to calculate the weight and similarity on each attribute.

2.2. Similarity Measurement for Categorical Attributes

For categorical attributes, each pair of values chosen from the value domain are considered to have the same distance as they do not have a natural ordering. By contrast, each pair of values of a numerical attribute has a numerical distance. Due to this different characteristic, it is not appropriate to use the Euclidian distance to evaluate categorical attributes-clustering similarity. Hereby, we adopt the frequency that the value $x^c_{i,r}$ appears in the cluster C_j for the categorical attribute A^c_r, where A^c_r $(r = 1, 2, \ldots, d_c)$ represents the rth categorical attribute.

Definition 1. *The similarity between a categorical attribute value $x^c_{i,r}$ and cluster C_j, where $i \in \{1, 2, \ldots, m\}$, $r \in \{1, 2, \ldots, d_c\}$, $j \in \{1, 2, \ldots, k\}$, is defined as*

$$s_c(x^c_{i,r}, C_j) = \frac{\sigma_{A^c_r = x^c_{i,r}}(C_j)}{\sigma_{A^c_r \neq NULL}(C_j)} \tag{3}$$

where $\sigma_{A^c_r = x^c_{i,r}}(C_j)$ represents the number of objects in cluster C_j, whose value for the categorical attribute A^c_r is equal to $x^c_{i,r}$, $NULL$ means empty, and $\sigma_{A^c_r \neq NULL}(C_j)$ represents the number of objects in cluster C_j, whose value for the categorical attribute A^c_r is not empty. From Definition 1, we can find the following properties:

1. $0 \leq s_c(x^c_{i,r}, C_j) \leq 1$;
2. $s_c(x^c_{i,r}, C_j) = 0$ only if none of the attribute A^c_r's values of the objects belonging to cluster C_j are equal to $x^c_{i,r}$;
3. $s_c(x^c_{i,r}, C_j) = 1$ only if all of the Non NULL attribute A^c_r's values of the objects belonging to cluster C_j are equal to $x^c_{i,r}$.

Optimizing attribute weights can improve the clustering performance. In information theory, the inhomogeneity degree of the dataset with respect to an attribute can be used to measure the significance of this attribute. In addition, according to Measure III proposed in [21], the higher the information content of an attribute, the higher the inhomogeneity degree of this attribute.

Definition 2. *Since the value domain of each attribute is definite, values of each attribute can be regarded as discrete and independent. The significance of an arbitrary categorical attribute A in dataset X can be quantified by the following entropy metric:*

$$H(A) = -\sum_{g=1}^{h} p(a_g) log p(a_g) \tag{4}$$

where A has a value domain, denoted as $dom(A)$, which consists of all the possible values that attributes A can choose, and $dom(A)$ can be represented with $dom(A) = \{a_1, a_2, \ldots, a_h\}$, h is the total number of values in $dom(A)$. $p(a_g) = \frac{\sigma_{A=a_g}(X)}{\sigma_{A \neq NULL}(X)}$, where a_g is a value of attribute A, $a_g \in dom(A)$, $g = 1, 2, \ldots, h$. Therefore, $p(a_g)$ is the probability density function of a_g in dataset X for attribute A. According to Equation (4), an attribute with more varying values has higher significance. However, in practice, an attribute with too many different values may have little clustering contribution, such as the instance ID number, which is unique for each instance; however, this information is useless for clustering analysis [12]. Thus, Equation (4) can be modified with Equation (5),

$$H(A^c_r) = -\frac{1}{h}\sum_{g=1}^{h} p(a_g) log p(a_g) \tag{5}$$

Then, the weight of each attribute based on information entropy is defined as in Equation (6),

$$w_r^c = \frac{H(A_r^c)}{\sum_{r=1}^{d_c} H(A_r^c) + \sum_{r=1}^{d_u} H(A_r^u)}, r = 1, 2, \ldots, d_c \tag{6}$$

where $\sum_{r=1}^{d_u} H(A_r^u)$ denotes the sum of modified information entropy of all the numeric attributes, which will be described in detail in the next section. Therefore, the metric function of similarity between object x_i^c and cluster C_j on categorical attributes is modified as Equation (7).

$$s_c(x_i^c, C_j) = \sum_{r=1}^{d_c} \frac{H(A_r^c)}{\sum_{r=1}^{d_c} H(A_r^c) + \sum_{r=1}^{d_u} H(A_r^u)} \frac{\sigma_{A_r^c = x_{i,r}^c}(C_j)}{\sigma_{A_r^c \neq NULL}(C_j)} \tag{7}$$

2.3. Similarity Measurement for Numerical Attributes

Since the entropy-based weighting strategy proposed in Section 2.2 is not applicable to numerical attributes, we made numerical data discrete at first. Then, the similarity measurement was used for the discretized data which are categorical data now. Discretization of numerical data is gaining more attention from the machine learning community [22]. Discretization of a given continuous attribute is also called quantization, which divides the range of attributes into intervals. Then, an interval label marks each interval. As a result, interval labels replace the original continuous data. Obviously, discretization can reduce the number of continuous attribute values [23], thereby simplifying the original data. Discretization also makes it possible for methods of categorical data clustering to be applied to cluster numerical or mixed datasets. There are many methods for numerical dataset discretization, such as discretization by intuitive division, histogram analysis, cluster analysis and entropy-based discretization. This section defines a smart way to automatically discretize numerical data by cluster analysis so that numerical data are transformed into categorical data. This method also provides a measure to find the optimal clusters to discretize the original data.

In order to transform numerical data into categorical data, we transformed numerical data by each attribute. Formally, let $X^l = [x_{1,l}^u, x_{2,l}^u, \ldots, x_{m,l}^u]$ be a single numerical attribute in X. X^l is transformed into the categorical values $\hat{X}_l = [\hat{x}_{1,l}, \hat{x}_{2,l}, \ldots, \hat{x}_{m,l}]$, $(l = 1, 2, \ldots, d_u)$. As a result, each point $x_i \in X (i = 1, 2, \ldots, m)$ is transformed into $\hat{x}_i = [x_{i,1}^c, x_{i,2}^c, \ldots, x_{i,d_c}^c, \hat{x}_{i,1}, \hat{x}_{i,2}, \ldots, \hat{x}_{i,d_u}]$.

Before categorizing numerical data by applying a clustering method to the data, the optimal number of categories is required, which is critical for the success of the categorization process.

Different methods have been proposed to find the optimal number of clusters for numerical attribute data [24,25]. The most common way is to apply a clustering algorithm to the data and calculate the cluster validity index. This process is repeated with an increasing number of clusters until it achieves the first local maxima. The number of clusters corresponding to the first local maxima is chosen as the optimal number of categories.

Let q be the number of clusters, which is unknown at first, and let f_q be a clustering function that assigns each $x_{i,l}^u \in X^l (i = 1, 2, \ldots, m)$ to one of the q clusters in Z^l, where $Z^l = \{z_1^l, z_2^l, \ldots, z_q^l\}$ and $z_j^l = \{x_{i,l}^u | f_q(x_{i,l}^u) \in z_j^l, i = 1, 2, \ldots, m, z_j^l \in Z^l\}$. The total sum of squares of X^l is defined as $S^l = \sum_{i=1}^m (x_{i,l}^u - \overline{x}^l)(x_{i,l}^u - \overline{x}^l)^T$, where $\overline{x}^l$ is the mean of X^l, which is defined as $\overline{x}^l = 1/m \sum_{i=1}^m x_{i,l}^u$. The within-cluster sum of squares is defined as $S_w^l(q) = \sum_{j=1}^q \sum_{x^l \in z_j^l} (x^l - u_j^l)(x^l - u_j^l)^T$, where the mean of each cluster u_j^l is defined as $u_j^l = \left(1 \middle/ |z_j^l|\right) \sum_{x^l \in z_j^l} x^l$. It can be found that $S_w^l(q)$ denotes the sum of deviations from each point to the center of their associated clusters, and the $S_w^l(q)$ of a good cluster should be a small value. The between-cluster sum of squares is defined as $S_b^l(q) = \sum_{j=1}^q |z_j^l|(u_j^l - \overline{x}^l)(u_j^l - \overline{x}^l)^T$, which denotes the sum of the weighted distances between each center of the q clusters and the center of data, and $S_b^l(q)$ of a good cluster result should be of a large value. It is clear that $S^l = S_w^l(q) + S_b^l(q)$; thus, the total sum of squares equals the sum of the within-cluster sum of squares and the between-cluster sum of squares.

The Calinski-Harabasz index is adopted to evaluate the clustering validity, which is defined as $S_{q,m}^l = \frac{(m-q)S_b^l(q)}{(q-1)S_w^l(q)}$. The proof of the effectiveness of the Calinski-Harabasz index is shown in [13]. We applied a clustering method f_q to the data X^l and calculated the corresponding Calinski-Harabasz index $S_{q,m}^l$ of clusters, $q = 2, 3, \ldots$. When the validity index $S_{q,m}^l$ achieved the first local maximum, we chose the corresponding q as the optimal number of categories, denoted as q_{best}^l.

To demonstrate the automatic categorization process, an example of the Calinski-Harabasz index calculation results is shown as Figure 1. In this example, the k-means method was chosen as f_q. When $q = 2, 3, \ldots, 100$, the k-means method was applied to the data and the validity index of the corresponding cluster result was calculated. When $q = 8$, the first local maxima of the Calinski-Harabasz index is found; therefore, $q_{best}^l = 8$. The automatic categorization process can be summarized as Algorithm 1.

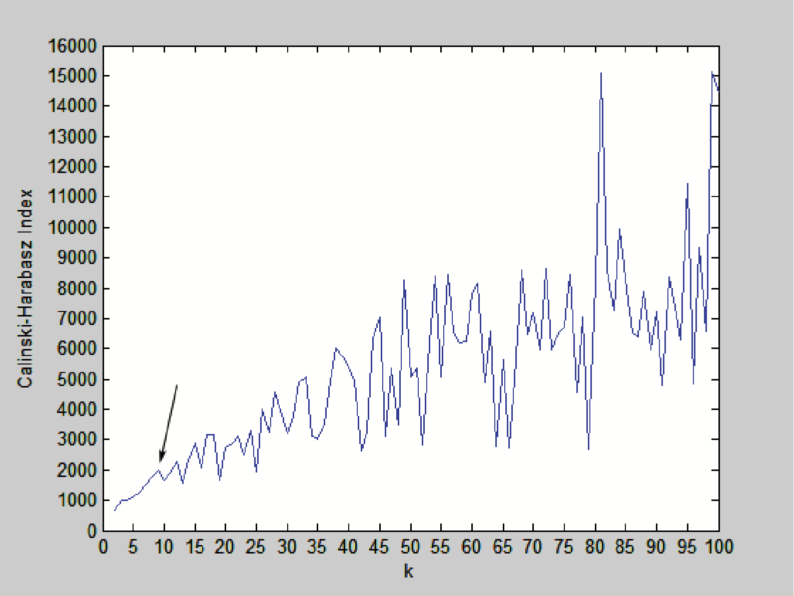

Figure 1. The Calinski-Harabasz index for $q = 2, 3, \ldots, 100$, the first local maxima of Calinski-Harabasz index is marked by the arrow.

Algorithm 1 Automatic categorization for numerical attributes.

```
Input: $X^l = [x_{1,l}^u, x_{2,l}^u, \ldots, x_{m,l}^u]$: the $l$th numerical attribute in the dataset $X$; $f_q$: a clustering function that partitions $X^l$ into $q$ clusters and returns the corresponding assignments; $q_{max}$: maxmum number of categories to examine;
Output: $\hat{X}^l = [\hat{x}_{1,l}^u, \hat{x}_{2,l}^u, \ldots, \hat{x}_{m,l}^u]$: the categorical values of $X^l$;
for $q = 2$ to $q_{max}$ do
    $S(q) = CalinskiHarabasz(f_q(X^l), X^l)$ (the Calinski-Harabasz index of the clustering result);
$q_{best} = min_{q \in \{2,3,\ldots,q_{max}\}}\{localMax(S(q)) = True\}$ (the first $q$ for which $S(q)$ achieves a local maxmum)
$\hat{X}^l = f_{q_{best}}(X^l)$
return $\hat{X}^l$
```

When the optimal number of categories q_{best}^l is found, each $x_{i,l}^u \in X^l, i = 1, 2, \ldots, m$ is allocated to one of q_{best}^l clusters $z_j^l \in Z^l$ by clustering method $f_{(}q_{best}^l)$, then the corresponding categorical value of $x^u_{i,l}$ is set as j. This process repeats for each numerical attribute X^l in dataset X, $l = 1, 2, \ldots, d_u$. After this automatic categorization process, we found the optimal number of categories of each numerical attribute and transformed the original numerical data into categorical data, with $\hat{x}^u_i = [\hat{x}^u_{i,1}, \hat{x}^u_{i,2}, \ldots, \hat{x}^u_{i,d_u}]$, $(i = 1, 2, \ldots, m)$.

Since the numerical data of the original dataset X is transformed into categorical data, we can use the similarity measurement for categorical data to the transformed data. The weight of each numerical attribute is calculated based on Equation (8):

$$w^u_r = \frac{H(A^u_r)}{\sum_{r=1}^{d_c} H(A^c_r) + \sum_{r=1}^{d_u} H(A^u_r)}, r = 1, 2, \ldots, d_u \tag{8}$$

where A^u_r, $(r = 1, 2, \ldots, d_u)$ represents each attribute of transformed data. Therefore, the metric function of similarity between object x^u_i and cluster C_j on numerical attribute is defined as:

$$\begin{aligned} s_u(x^u_i, C_j) &= \sum_{r=1}^{d_u} w^u_r s_u(x^u_{i,r}, C_j) \\ &= \sum_{r=1}^{d_u} w^u_r s_u(\hat{x}^u_{i,r}, C_j) \\ &= \sum_{r=1}^{d_u} w^u_r \frac{\sigma_{A^u_r = \hat{x}^u_{i,r}}(C_j)}{\sigma_{A^u_r \neq NULL}(C_j)} \end{aligned} \tag{9}$$

2.4. Similarity Measurement for Mixed Data

Combining Sections 2.2 and 2.3, the similarity measurement for mixed data is defined as:

$$\begin{aligned} s(x_i, C_j) =& \sum_{r=1}^{d_c} w^c_r s_c(x^c_{i,r}, C_j) + \sum_{r=1}^{d_u} w^u_r s_u(x^u_{i,r}, C_j) \\ =& \sum_{r=1}^{d_c} w^c_r s_c(x^c_{i,r}, C_j) + \sum_{r=1}^{d_u} w^u_r s_u(\hat{x}^u_{i,r}, C_j) \\ =& \sum_{r=1}^{d_c} w^c_r \frac{\sigma_{A^c_r = x^c_{i,r}}(C_j)}{\sigma_{A^c_r \neq NULL}(C_j)} + \sum_{r=1}^{d_u} w^u_r \frac{\sigma_{A^u_r = \hat{x}^u_{i,r}}(C_j)}{\sigma_{A^u_r \neq NULL}(C_j)} \end{aligned} \tag{10}$$

2.5. Iterative Clustering Algorithm

Based on Equation (10), the similarity measurement with entropy-based weighting applied to the k-means framework can be conducted as Algorithm 2.

Steps 1–3 utilize the automatic categorization process to obtain transformed categorical datasets based on Algorithm 1. Since the attributes of the transformed dataset are all categorical, the weight of each attribute can be calculated with the entropy-based weighting strategy, and Steps 4–9 show the process. Steps 10–21 are the iterative process that applies the similarity measurement based on Equation (10) into the k-means algorithm framework to address the transformed dataset.

Algorithm 2 Iterative clustering algorithm with entropy-based weighting.

```
Input: $X = \{x_1, x_2, \ldots, x_m\}$ (dataset to cluster with $d_c$ categorical attributes and $d_u$ numerical attributes);
    $k$ (number of clusters);
    $f_q$ (a clustering function that partitions $X^l$ into $q$ clusters and returns the corresponding assignments);
    $q_{max}$ (maximum number of categories to examine);
Output: $idx = \{idx_1, idx_2, \ldots, idx_m\}$ (an assignment of each point in $X$ to one of $k$ clusters);
for $l = 1$ to $d_u$ do
    $\hat{X}^l = Categorize(X^l, f_q, q_{max})$ (automatic categorization of $X^l$);
for $r = 1$ to $d_c$ do
    $w_r^c = \frac{H(A_r^c)}{\sum_{r=1}^{d_c} H(A_r^c) + \sum_{r=1}^{d_u} H(A_r^u)}$ (calculate the importance of each categorical attribute);
for $r = 1$ to $d_u$ do
    $w_r^u = \frac{H(A_r^u)}{\sum_{r=1}^{d_c} H(A_r^c) + \sum_{r=1}^{d_u} H(A_r^u)}$ (calculate the importance of each numerical attribute);
Set $idx = \{0, 0, \ldots, 0\}$ and select $k$ initial objects randomly as $k$ initial centroids for each cluster
$noChange = true$;
repeat
    for $i = 1$ to $m$ do
        $idx_i^{(new)} = argmax_{j \in \{1,2,\ldots,k\}}[s(x_i, C_j)]$;
        if $idx_i^{(new)} \neq idx_i^{(old)}$ then
            $noChange = false$;
            Update the information of clusters $C_{idx_i}^{(new)}$ and $C_{idx_i}^{(old)}$, including the frequency of each categorical value.
until ($noChange = true$)
return $idx$
```

3. Results and Discussion

To test the effectiveness of the similarity measurement with the entropy-based weighting proposed in this paper, two different types of datasets, mixed and numerical datasets, were selected from the UCI Machine Learning Data Repository [26], and most datasets were collected from the field of biology and medicine. The iterative clustering algorithm based on the proposed similarity measurement was compared with existing clustering algorithms, including OCIL [12], K-Prototype [9] and k-means [4]. k-means was used for dataset made of numerical variables only. In the experiments, the clustering accuracy [27] was adopted to evaluate the three mentioned methods. The clustering accuracy is defined as $AC = \frac{\sum_{i=1}^{m} \delta(l_i, map(idx_i))}{m}$, where m denotes the number of instances of the dataset, l_i denotes the provided label, idx_i denotes the obtained cluster label, $map(idx_i)$ is a mapping function that maps idx_i to the equivalent label from the data corpus, and the function $\delta(l_i, map(idx_i)) = 1$ only if $l_i = map(idx_i)$; otherwise, the value is 0. Correspondingly, the clustering error rate is defined as $error = 1 - AC$.

In the experiments, considering that the clustering results are affected by the selected initial centroids, we set the same initial centroids for all methods during each test, and the following experimental results were averaged from 100 random runs. In addition, k-means was chosen as the clustering method in the automatic categorization process that transforms numerical attributes into categorical attributes. Before using k-means method to cluster, the original data were normalized to be between 0 and 1.

3.1. Experiments on Mixed Datasets

In this section, we investigated the performance of the iterative clustering algorithm based on the proposed similarity measurement on mixed datasets. Table 1 shows the information of each dataset. Note that the second column presents the number of samples, the third column presents the number of the two types of attributes, and the last column presents the probability distribution of samples in different classes.

To evaluate the performance of the iterative clustering algorithm based on the proposed similarity measurement, we compare its clustering results with OCIL and K-Prototype. The average value and standard deviation of the clustering error of these clustering algorithms are statistically summarized in Table 2. In the experiments, the weight parameter γ was set to 1.5 for the K-Prototypes algorithm.

Table 1. Description of mixed datasets.

Dataset	Instance	Attribute ($d_c + d_u$)	Class	Class Distribution
Statlog(Heart)	270	7 + 6	2	55.56%, 44.44%
Hepatitis	155	13 + 6	2	20.65%, 79.35%
Cylinder Bands	540	19 + 20	2	42.33%, 57.67%
Australian	690	8 + 6	2	55.51%, 44.49%
Dermatology	366	33 + 1	6	30.60%, 16.67%, 19.67%, 13.39%, 14.21%, 5.46%
Zoo	101	16 + 1	7	40.59%, 19.80%, 4.95%, 12.87%, 3.96%, 7.92%, 9.90%

Australian denotes Statlog (Australian Credit Approval) Dataset.

Table 2. Comparison of cluster accuracy for the proposed algorithm with OCIL and K-Prototype on mixed datasets.

Dataset	OCIL	K-Prototype	The Proposed Algorithm
Statlog (Heart)	0.1891 ± 0.0029	0.2192 ± 0.0662	0.1606 ± 0.0288
Hepatitis	0.2065 ± 0.0000	0.2065 ± 0.0000	0.1810 ± 0.0074
Cylinder Bands	0.2743 ± 0.0861	0.2852 ± 0.0749	0.2676 ± 0.1168
Australian	0.2579 ± 0.1286	0.2218 ± 0.0678	0.2136 ± 0.0686
Dermatology	0.1953 ± 0.0510	0.3063 ± 0.0792	0.1855 ± 0.0555
Zoo	0.1449 ± 0.0376	0.1578 ± 0.0520	0.1318 ± 0.0341

From Table 2, it can be observed that the iterative clustering algorithm based on the proposed similarity measurement outperforms the OCIL and K-Prototype methods for six datasets, although the ratios of the numbers of categorical attributes to numerical attributes differ greatly, as shown in Table 1. Compared to the other two methods, the iterative clustering algorithm can improve the accuracies of clustering results by 2.13% and 4.28%, respectively. Especially in the Heart dataset, the iterative clustering algorithm improves the accuracy by 2.85% and 5.86%, respectively. This result indicates that the proposed similarity measurement is applicable to mixed datasets of variant compound styles and does not need any parameter to give weights to the two types of attributes. Furthermore, for datasets that have very uneven class distributions, the proposed similarity measurement can also achieve adequate clustering results.

To study why the iterative clustering algorithm based on the proposed similarity measurement outperforms the OCIL and K-Prototype methods, we analyzed the correlation between each attribute and the label attribute in the dataset by calculating the correlation coefficients in statistics. Since the label attribute is a categorical attribute, the Pearson correlation coefficient and the Spearman correlation coefficients are not suitable. The Kendall correlation coefficient requires that the categorical attribute be ordered; therefore, it also cannot be used to calculate the correlation between the categorical attribute and the label attribute. Here, we adopt the ReliefF algorithm, which can estimate the quality of dependencies between each attribute and label attribute [28].

To see whether the correlation between the dependencies and weights affects the clustering results, we calculate the Pearson correlation coefficient between the dependencies calculated by the ReliefF algorithm and the weights calculated by Equations (6) and (8) in each dataset. The result is shown in Table 3. Since the Dermatology dataset as well as the

Zoo dataset have only one numerical attribute, calculating the correlation for numerical attribute is meaningless.

Table 3. The correlation coefficient between dependencies and weights in each mixed dataset.

Dataset	Correlation Coefficient for Categorical Attributes	Correlation Coefficient for Numerical Attributes
Statlog (Heart)	−0.2602	0.3885
Hepatitis	−0.0151	0.7953
Cylinder Bands	0.3648	0.3955
Australian	0.9314	0.7281
Dermatology	0.7822	\
Zoo	−0.2878	\

Comparing Tables 2 and 3, it can be seen that the Dermatology dataset, with a good clustering result, has a strong correlation between dependencies and weights for categorical attributes. In addition, the Australian dataset has a strong correlation not only for categorical attributes but also for numerical attributes; this dataset also has a good clustering result. However, in the Hepatitis dataset it has a strong correlation for categorical attributes, but has a weak correlation for numerical attributes. There are only six numerical attributes and 13 categorical attributes in the Hepatitis dataset. The influence of category attributes is much greater than that of numerical attributes. So, the clustering result of Hepatitis dataset is not good, which does not violate the theory in the article. Therefore, a good clustering result may be obtained due to the reasonable weight assigned to each attribute by the proposed similarity measurement.

3.2. Experiments on Numerical Datasets

Then, we further investigated the performance of the proposed similarity measurement on pure numerical datasets. Table 4 shows the information of six numerical datasets, including the number of samples, attributes and classes, and class distribution. To evaluate the performance of the proposed similarity measurement applied to k-means on numerical datasets, we also conducted experiments compared to the most classical numerical data clustering algorithms, k-means algorithms. These two clustering algorithms were applied to different numerical datasets; Table 5 shows the mean and variance of the clustering error of clustering by applying different algorithms. In addition, before using the k-means method to cluster, the dataset was normalized to between 0 and 1.

Table 4. Description of numerical datasets.

Dataset	Instance	Attribute	Class	Class Distribution
Waveform	5000	40	3	33.84%, 33.06%, 33.10%
Wine	178	13	3	33.15%, 39.89%, 26.97%
Mass	961	5	2	53.69%, 46.31%
Seeds	210	7	3	33.33%, 33.33%, 33.33%
Iris	150	4	3	33.33%, 33.33%, 33.33%
Fertility	100	9	2	88.00%, 12.00%

Waveform denotes waveform-+noise Dataset in Waveform Database Generator (Version 1)
Mass denotes Mammographic Mass Dataset.

Table 5. Comparison of the proposed algorithm with k-means on numerical datasets.

Dataset	k-Means	The Proposed Algorithm
Waveform	0.4764 ± 0.0002	0.4654 ± 0.0097
Wine	0.0378 ± 0.0031	0.0660 ± 0.0451
Mass	0.4631 ± 0.0000	0.1962 ± 0.0740
Seeds	0.3857 ± 0.0000	0.3813 ± 0.0310
Iris	0.1677 ± 0.0859	0.0563 ± 0.0651
Fertility	0.1200 ± 0.0000	0.1200 ± 0.0000

It can be seen that except for the Wine dataset and the Fertility dataset, the proposed similarity measurement applied to k-means outperforms the k-means method on other datasets. Normally, the clustering accuracy of the iterative clustering algorithm based on the proposed similarity measurement is 6.09% higher than that of k-means, especially for the Mass dataset, with a similarity that is 26.69% higher than that of k-means.

Similarly, the correlation coefficient between dependencies and weights was calculated to analyze why the iterative clustering algorithm outperforms the k-means method, and the correlation coefficient of each dataset is shown in Table 6. It can be found that the Mass dataset with the best clustering result has the highest correlation coefficient. Perhaps the weights of this dataset are well allocated according to the contribution of each clustering attribute.

Table 6. The correlation coefficient between dependencies and weights in each numerical dataset.

Dataset	Correlation Coefficient
Waveform	0.1200
Wine	−0.1458
Mass	0.9245
Seeds	0.0944
Iris	−0.6511
Fertility	0.5526

4. Conclusions

In this paper, a similarity measurement with entropy-based weighting is proposed for mixed datasets with numerical and categorical attributes. For categorical datasets, a similarity metric is designed by assigning different weights to each attribute based on information entropy theory. For numerical datasets, the original high-dimensional numerical data are transformed to categorical data by an automatic categorization technique, so the similarity metric for categorical datasets can be applied to numerical datasets. Then, a similarity measurement for mixed datasets was obtained. Extensive experimental results show that the iterative clustering algorithm based on the proposed similarity measurement can achieve higher clustering accuracy and is superior to the existing clustering algorithms on datasets from UCI. The results also validate the feasibility of handling different types of attributes and verify that various attributes contribute differently in similarity measurements when clustering.

Author Contributions: Conceptualization, X.Q. and S.J.; methodology, J.Y.; software, S.J.; writing—original draft preparation, J.Y. and S.J.; writing—review and editing, J.Y. and X.Q.; supervision, N.A. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported partially by the National Natural Science Foundation of China (No. 61502135), the Programme of Introducing Talents of Discipline to Universities (No. B14025), and the Anhui Provincial Key Technologies R&D Program (No. 1804b06020378). The funding bodies had no role in study design, data collection and analysis, or preparation of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jiawei, H.; Micheline, K. Data Mining: Concepts and Techniques. *Data Min. Concepts Model. Methods Algorithms Second Ed.* **2006**, *5*, 1–18.
2. Rodoshi, R.T.; Kim, T.; Choi, W. Resource Management in Cloud Radio Access Network: Conventional and New Approaches. *Sensors* **2020**, *20*, 2708. [CrossRef]
3. Khorraminezhad, L.; Leclercq, M.; Droit, A.; Bilodeau, J.F.; Rudkowska, I. Statistical and Machine-Learning Analyses in Nutritional Genomics Studies. *Nutrients* **2020**, *12*, 3140. [CrossRef] [PubMed]
4. Macqueen, J. Some Methods for Classification and Analysis of Multivariate Observations. *Berkeley Symp. Math. Stat. Probab.* **1967**, *1*, 281–297.
5. Ahmad, A.; Hashmi, S. K-Harmonic means type clustering algorithm for mixed datasets. *Appl. Soft Comput.* **2016**, *48*, 39–49. [CrossRef]
6. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. R. Stat. Soc.* **1977**, *39*, 1–38.
7. Cao, F.; Liang, J.; Li, D.; Bai, L.; Dang, C. A dissimilarity measure for the k-Modes clustering algorithm. *Knowl. Based Syst.* **2012**, *26*, 120–127. [CrossRef]
8. Guha, S.; Rastogi, R.; Shim, K. ROCK: A robust clustering algorithm for categorical attributes. *Inf. Syst.* **1999**, *25*, 345–366. [CrossRef]
9. Huang, Z. Clustering large data sets with mixed numeric and categorical values. In Proceedings of the 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining, Singapore, 23–24 February 1997; pp. 21–34.
10. Ahmad, A.; Khan, S. Survey of State-of-the-Art Mixed Data Clustering Algorithms. *IEEE Access* **2019**, *7*, 31883–31902. [CrossRef]
11. Huang, Z. Extensions to the k-means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Min. Knowl. Discov.* **1998**, *2*, 283–304. [CrossRef]
12. Cheung, Y.M.; Jia, H. Categorical-and-numerical-attribute data clustering based on a unified similarity metric without knowing cluster number. *Pattern Recognit.* **2013**, *45*, 2228–2238. [CrossRef]
13. David, G.; Averbuch, A. SpectralCAT: Categorical spectral clustering of numerical and nominal data. *Pattern Recognit.* **2012**, *45*, 416–433. [CrossRef]
14. Ng, A.; Jordan, M.; Weiss, Y. On spectral clustering: Analysis and an algorithm. *Adv. Neural Inf. Process. Syst.* **2001**, *14* 849–856.
15. Hsu, C.C. Generalizing self-organizing map for categorical data. *IEEE Trans. Neural Netw.* **2006**, *17*, 294–304. [CrossRef]
16. Liang, J.; Chin, K.S.; Dang, C.; Yam, R.C. A new method for measuring uncertainty and fuzziness in rough set theory. *Int. J. Gen. Syst.* **2002**, *31*, 331–342. [CrossRef]
17. Ng, M.K.; Li, M.J.; Huang, J.Z.; He, Z. On the impact of dissimilarity measure in k-modes clustering algorithm. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 503. [CrossRef]
18. Chen, L.F.; Guo, G.D. Non-mode clustering of categorical data with attributes weighting. *J. Softw.* **2013**, *14*, 2628–2641. [CrossRef]
19. Bai, L.; Liang, J.; Dang, C.; Cao, F. A novel attribute weighting algorithm for clustering high-dimensional categorical data. *Pattern Recognit.* **2011**, *44*, 2843–2861. [CrossRef]
20. Ahmad, A.; Dey, L. A k-mean clustering algorithm for mixed numeric and categorical data. *Data Knowl. Eng.* **2007**, *63*, 503–527. [CrossRef]
21. Basak, J.; Krishnapuram, R. Interpretable Hierarchical Clustering by Constructing an Unsupervised Decision Tree. *IEEE Trans. Knowl. Data Eng.* **2005**, *17*, 121–132. [CrossRef]
22. Dougherty, J.; Kohavi, R.; Sahami, M. Supervised and Unsupervised Discretization of Continuous Features. *Mach. Learn. Proc.* **1995**, *2*, 194–202.
23. Grzymala-Busse, J.W. Data reduction: Discretization of numerical attributes. *Handbook of Data Mining and Knowledge Discovery*; Oxford University Press, Inc.: Oxford, UK, 2002; pp. 218–225.
24. Jung, Y.; Park, H.; Du, D.Z.; Drake, B.L. A Decision Criterion for the Optimal Number of Clusters in Hierarchical Clustering. *J. Glob. Optim.* **2003**, *25*, 91–111. [CrossRef]
25. Bayati, H.; Davoudi, H.; Fatemizadeh, E. A heuristic method for finding the optimal number of clusters with application in medical data. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* **2008**, *2008*, 4684–4687.
26. UCI Machine Learning Repository. Available online: http://archive.ics.uci.edu/ml (accessed on 15 June 2021).
27. Zhu, L.; Miao, L.; Zhang, D. Iterative Laplacian Score for Feature Selection. In *Chinese Conference on Pattern Recognition*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 80–87.
28. Kononenko, I. Estimating attributes: Analysis and extensions of RELIEF. In *European Conference on Machine Learning*; Springer: Berlin/Heidelberg, Germany, 1994; pp. 171–182.