A Novel Three-way Clustering Algorithm for Mixed-type Data

Hong Yu*, Zhihua Chang*, and Bing Zhou†
*Chongqing Key Laboratory of Computational Intelligence
Chongqing University of Posts and Telecommunications, Chongqing, 400065, China
Email: yuhong@cqupt.edu.cn, zhihua_chang@qq.com
†Department of Computer Science, Sam Houston State University, Huntsville, Texas, 77341, USA
Email: bxz003@SHSU.EDU

***Abstract*—Large quantities of mixed-type data, containing categorical, ordinal and numerical attributes, have commonly existed in real world. In this paper, a mixed-type data clustering method, which could deal with the uncertain situation that a cluster may not have a definite cluster boundary, is proposed inspired by the theory of three-way decisions. Many existing studies represent a cluster with a single set based on a two-way strategy, which does not adequately show the fact that a cluster may not have a well-defined cluster boundary. In this paper, we represent a cluster with a pair of sets, i.e., the core region and fringe region. The three-way clustering is suitable for dealing with uncertainty because it shows intuitively which objects are fringe to the cluster. Then, new measurements of distance between mixed-type data are proposed for different types of attribute values by means of a weighted tree structure. The measurement considers the semantic of attributes, the number of attribute values and the occurrence frequency of attribute values. Finally, a three-way clustering algorithm for mixed-type data is proposed. The experimental results show that the proposed distance measure of mixed-type data is reasonable and effective, the proposed algorithm is in a better performance at the accuracy and the average adjusted rand index than the compared algorithms in most cases.**

***Keywords*-cluster; mixed-type data; three-way decisions; uncertainty;**

I. Introduction

Cluster analysis is one of the most important techniques used in data mining and machine learning for grouping, so that objects in the same group are more similar to each other than to those outside the group [**?**], [**?**]. The data is characterized by numerical attributes, or characterized by categorical attributes. Even in some cases, there exists ordinal relationship between attribute values.

In order to solve the problem of mixed-type data clustering, Huang [**?**] proposed the famous K-prototypes algorithm that combined K-means with K-modes algorithm. Then, some derivations of K-prototypes algorithm have been presented [**?**], [**?**], [**?**]. Pathak and Pal [**?**] used collaborative clustering method to find the cluster substructures that are common to both categorical and numerical part of the mixed data, then obtained the final clustering result. Chen and He [**?**] thought that there are three types of mixed-type data such as the numerical dominant data, the categorical dominant data and the balanced data, and they designed the corresponding measurement function. Lam et al. [**?**] used the fuzzy adaptive resonance theory to cluster the mixed-type data with categorical and numeric features. Aziz [**?**] proposed a new dissimilarity measure based on eigenstructure of the covariance matrix and robust principal component score to achieve mixed data clustering.

Therefore, for further improving the similarity measure of mixed-type data containing categorical data and numerical data as well as ordinal data, Hsu et al. [**?**], [**?**] constructed different tree structures for categorical data and ordinal data, and the hierarchical distance between attribute values in the tree structure is used to calculate the distance of data. However, it is not a convincing method to assign the weights randomly in advance. Aiming at computing the categorical data, more and more scholars [**?**], [**?**], [**?**] think that the number and the occurrence frequency of attribute values have important influence on similarity measure. Therefore, it is very important to find an effective measurement which considers semantics of attribute values and the characteristics of data sets.

On the other hand, many existing clustering approaches represent a cluster by a single set based on a two-way strategy. That is, one object belongs to a cluster or not belong to a cluster. However, in the actual production, such as in pattern recognition system [**?**], there is not only a situation that one object definitely belongs to or not belong to a cluster, but also a situation that one object might or might not belong to a cluster. The traditional crisp two-way representation does not adequately show the fact that a cluster may not have a well-defined cluster boundary. Thus, we proposed the framework of three-way clustering [**?**], [**?**], [**?**] inspired by the three-way decisions, which is introduced by Professor Yao [**?**], [**?**]. Three-way decisions extend binary-decisions in order to overcome some drawbacks of binary-decisions. The basic ideas of three-way decisions have been widely used in real-world decision-making problems, such as clustering analysis, decision making, email spam filtering, three-way investment decisions and many others [**?**].

In the three-way clustering, a cluster is represented by a pair of sets which divide a cluster into three regions, i.e., the core region, fringe region and trivial region, instead of two regions produced by a single set as the other existing

978-1-5386-3120-1/17 $31.00 © 2017 IEEE
DOI 10.1109/ICBK.2017.38

methods. Objects in the core region are typical elements of the cluster, objects in the fringe region are fringe elements of the cluster, and objects in the trivial region do not belong to the cluster definitely. A cluster is therefore more realistically characterized by a set of core objects and a set of boundary objects. The three-way representation intuitively shows which objects are fringe to the cluster. Obviously, the classic two-way representation is a special case of three-way representation when the fringe region is empty.

In this paper, for different types of attribute values, we construct different measurement of distance between objects, which is based on a weighted tree structure through considering the influence of the semantics of attribute values, the number of attribute values and the occurrence frequency of the attribute values. Then, a novel three-way clustering algorithm for mixed-type data is proposed.

The remainder of this paper is organized as follows. Section **??** introduces some basic definitions such as mixed-type data and the representation of three-way clustering. Section **??** proposes a novel three-way decision clustering framework for mixed-type data. Section **??** reports the results of comparative experiments and conclusions are provided in Section **??**.

II. Preliminaries

A. Mixed-type data

Let us consider a universe $\mathbf{U} = \{\mathbf{x_1}, \cdots, \mathbf{x_n}, \cdots, \mathbf{x_N}\}$ with N objects. A D-dimensional data point $\mathbf{x_n}$ comprises P categorical attributes, Q numerical attributes and T ordinal attributes, where $P + Q + T = D$, i.e., $\mathbf{x_n} = \{x_n^1, \cdots, x_n^P, x_n^{P+1}, \cdots, x_n^{P+Q}, x_n^{P+Q+1}, \cdots, x_n^D\}$.

$\left\{x_n^1, x_n^2, \cdots, x_n^P\right\}$ is a subset of categorical attributes without sequence meaning, in other words, there is no order or rank relationship between these attribute values. The attribute values represent a certain class, coding, state, etc. $\left\{x_n^{P+1}, x_n^{P+2}, \cdots, x_n^{P+Q}\right\}$ represents a subset of ordinal attributes which attribute values have meaningful order or rank. The attribute values are representative by the sequence of words, symbols, numbers, etc. For example, the attribute values have the ordinal meaning in the elements of the set $\{Very\ bad, Bad, Medium, Good, Very\ good\}$. $\left\{x_n^{P+Q+1}, x_n^{P+Q+2}, \cdots, x_n^{P+Q+T}\right\}$ is a set of numerical-valued attributes.

B. Representation of three-way clustering

The purpose of clustering is to divide objects in a universe into some clusters. If there are K clusters, the family of clusters, $\mathbf{C}$, is represented as $\mathbf{C} = \{C_1, \cdots, C_k, \cdots, C_K\}$. In many existing clustering approaches, a cluster is represented by a single set, where the objects in the set belong to this cluster definitely and the objects not in the set do not belong to this cluster definitely. This is a typical result of two-way decisions. However, this representation is not adequately show the fact that a cluster may not have a well-defined cluster boundary, especially in real applications. Thus, the representation of three-way clustering using a pair of sets to represent a cluster is more appropriate than the use of a crisp set, which also directly leads to three-way decisions based interpretation of clustering.

In contrast to the general crisp representation of a cluster, we represent a three-way cluster C_k as a pair of sets:

$$C_k = \left(Co\left(C_k\right), Fr\left(C_k\right)\right), \tag{1}$$

where $Co\left(C_k\right) \subseteq \mathbf{U}$ and $Fr\left(C_k\right) \subseteq \mathbf{U}$. let $Tr\left(C_k\right) = \mathbf{U} - Co\left(C_k\right) - Fr\left(C_k\right)$.

If $Fr(C_k) = \emptyset$, the representation of C_k in Eq. (**??**) turns into $C_k = Co(C_k)$; it is a single set and $Tr(C_k) = U - Co(C_k)$. This is a representation of two-way decisions. In other words, the representation of a single set is a special case of the representation of three-way cluster.

$Co\left(C_k\right)$, $Fr\left(C_k\right)$, $Tr\left(C_k\right)$ naturally form the three regions of a cluster as core region, fringe region and trivial region respectively. Objects in the core region of a cluster definitely belong to the cluster, objects in the trivial region of a cluster definitely do not belong to the cluster C_k, and objects in the fringe region of a cluster might or might not belong to the cluster.

These subsets have the following properties.

$$\begin{aligned} &Co\left(C_k\right) \cap Fr\left(C_k\right) = \phi, \\ &Co\left(C_k\right) \cap Tr\left(C_k\right) = \phi, \\ &Fr\left(C_k\right) \cap Tr\left(C_k\right) = \phi. \end{aligned} \tag{2}$$

Furthermore, according to Formula (**??**), we know that it is enough to represent a cluster expediently by the core region and the fringe region.

In another way, we can define a cluster by the following properties:

$$\begin{aligned} &(i) Co\left(C_k\right) \neq \phi, 0 < k < K, \\ &(ii) Co\left(C_k\right) \cup Fr\left(C_k\right) \cup Tr\left(C_k\right) = \mathbf{U}. \end{aligned} \tag{3}$$

Property (i) implies that a cluster cannot be empty. This makes sure that a cluster is physically meaningful. And property (ii) states that any object in $\mathbf{U}$ must definitely belong to or might belong to a cluster, which ensures that every object is properly clustered.

With respect to the family of clusters, $\mathbf{C}$, we have the following family of clusters formulated by three-way decisions as:

$$\mathbf{C} = \left\{\left(Co\left(C_1\right), Fr\left(C_1\right)\right), \cdots, \left(Co\left(C_K\right), Fr\left(C_K\right)\right)\right\}. \tag{4}$$

Under the representation, we can re-formulate the clustering as follows. For a clustering, if there exists $k \neq t$, such that

$$\begin{aligned} &(1)\ Co\left(C_k\right) \cap Co\left(C_t\right) \neq \phi, or \\ &(2)\ Fr\left(C_k\right) \cap Fr\left(C_t\right) \neq \phi, or \\ &(3)\ Co\left(C_k\right) \cap Fr\left(C_t\right) \neq \phi, or \\ &(4)\ Fr\left(C_k\right) \cap Co\left(C_t\right) \neq \phi. \end{aligned} \tag{5}$$

As long as one condition is satisfied, it is called an overlapping (or soft) clustering; otherwise, it is a hard clustering. As long as one condition from Eq. (**??**) is satisfied, there must exist at least one object belonging to more than one cluster.

III. The Proposed Method

In this section, we first introduce the measurement of distance between objects. The different weighted tree are utilized to represent and measure the distance between categorical values, ordinal values and numerical values, respectively. Then, a novel three-way clustering algorithm for mixed-type data is presented.

A. The measurement of distance

1) Categorical values: The proposed distance measurement borrows the representation of distance hierarchy [**?**]. Actually, the distance hierarchy structure is a kind of tree in topological structure. The tree consists of nodes, edges, and weights as shown in Figure **??**. We need to construct a tree for every categorical attribute. The leaf nodes are composed of the attribute values. The distance between any two categorical values is the distance between the two leaf nodes in the weighted tree. The computing process concludes three steps: 1) to establish the topological structure according to the semantic of attribute values; 2) to weight the edges according to the number of attribute values, the occurrence frequency of attribute values and the depths of the node; and 3) to compute the distance between leaf nodes on the tree.

Let Cat^d be the d-th categorical attribute in the original data set $\mathbf{U}$, and $V(Cat^d)$ be the domain of Cat^d. There are three principles to construct a weighted tree. First, one leaf node represents one attribute value of Cat^d; that is to say, the number of leaf nodes is equal to the cardinality of $V(Cat^d)$. Second, if the hierarchy of semantic is deeper, the corresponding subtree is closer to the right side. Third, on the same hierarchy, higher the frequency of values is, the corresponding node is closer to the right side.

To explain the tree, we illustrate an example here. There is a categorical attribute, i.e., $Drink$, and it has 8 attribute values, that is $V(Drink)$ = {boiled water, pepsi, coke, sprite, strawberry juice, orange juice, green apple juice, red apple juice}. The semantic of values tell us that {pepsi, coke, sprite} belong to "carbonated" drinks, {strawberry juice, orange juice, green apple juice, red apple juice} belong to "juice", and {green apple juice, red apple juice} are "apple juice". Then, we get the topological structure as shown in Figure **??**. The $root^d$ is $Drink$, the internal nodes are the categorical classed and the leaf nodes are the eight values. The digit under of the leaf nodes is the frequency of the value in the attribute column. For example, the value "red apple" appears 40 times.

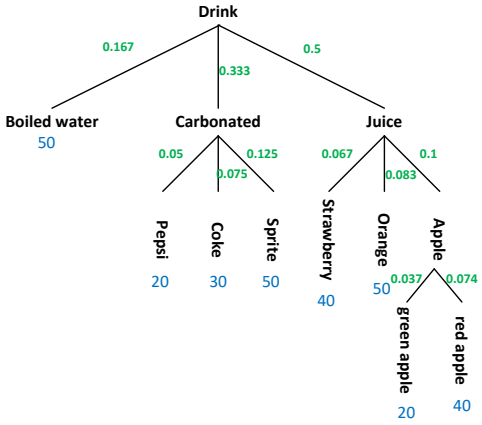

Figure 1. The weighted tree of categorical values

Let $node_i^h$ represents the i-th node at the h level. For example, in Figure **??**, $node_3^2$ means the node $juice$. Let $child\left(node_i^h\right)$ be a set of child nodes of $node_i^h$, and $child_j\left(node_i^h\right)$ means its the j-th child node, $\left|child\left(node_i^h\right)\right|$ is the number of child nodes of $node_i^h$. For example, $\left|child\left(node_3^2\right)\right| = 3$, it means node $juice$ has three child nodes.

F. Noorbehbahani [**?**] pointed out that for any categorical attribute, the more the number of attribute values is, the smaller the distance between attribute values is. For example, if a categorical attribute is supposed to have a domain size of 2, the distance of two unmatched attribute values will be greater than the case having domain size of 20. We use the weight on the edge of the parent node and the child node to represent the distance between them. Then, the number of children reflects the values of weight. Obviously, the more the number of child nodes is, the smaller the weight between parent node and the child node is.

Thus, derived from [**?**], we first define the weight function $f(node_i^h)$ as follows. The function compute the weight coefficient in view of influence by the number of children on an internal node $(node_i^h)$.

$$f(node_i^h) = \begin{cases} 1, & x \le \theta_1 \\ 1 - \xi_1 (x - \theta_1), & \theta_1 < x \le \theta_2 \\ 1 - \xi_1 (\theta_2 - \theta_1) - \xi_2 (x - \theta_2), & x > \theta_2 \end{cases} \tag{6}$$

where $x = \left|child\left(node_i^h\right)\right|$, $\xi_1, \xi_2 \in (0,1)$, $\theta_1, \theta_2 \in N+$.

According to Eq. (**??**), we know that with the growth of the number of child nodes, the weight coefficient between parent node and child node is getting smaller.

Meanwhile, according to the conclusions in [**?**], [**?**], [**?**], we know that for categorical attribute Cat^k, the partition ability for clustering is affected by the occurrence frequency of its attribute values in the data set.

Let $fr\left(node_i^h\right)$ represent its occurrence frequency in a data set. When it is a leaf node, $fr\left(node_i^h\right)$ is the frequency of the value. When it is an internal node, $fr\left(node_i^h\right)$ is

the sum of the occurrence frequency of its all children. In other words, $fr\left(node_i^h\right)$ is the occurrence frequency in the subtree whose root is the node $node_i^h$. For example, in Figure **??**, $fr\left(carbonated\right) = fr\left(pepsi\right) + fr\left(coke\right) + fr\left(sprite\right)$=100.

On the other hand, we find that the weight between parent node and child node is relevant to the levels (depths) where they are in the tree. For example, in Figure **??**, although "strawberry juice", "orange juice", "green apple juice", "red apple juice" are all belong to "juice", the distance between "strawberry juice" and "orange juice" should be bigger than the distance between "green apple juice" and "red apple juice". Because "green apple juice" and "red apple juice" are both belong to "apple juice" while "strawberry juice" and "orange juice" are just belong to "juice". In other words, "green apple juice" and "red apple juice" are more similar than "strawberry juice" and "orange juice".

Based on the above analysis, the influence factors in weight between parent node and its child node contain the number of child nodes, the frequency and the level of nodes.

Let $w(node_i^h, child_j(node_i^h))$ be the weight between $node_i^h$ and its j-th child node $child_j\left(node_i^h\right)$. It is calculated as follows.

$$w(node_i^h, child_j(node_i^h)) = \frac{f\left(node_i^h\right) \cdot \frac{fr\left(child_j\left(node_i^h\right)\right)}{fr\left(node_i^h\right)}}{h^\eta}. \tag{7}$$

where $\eta \in N+$.

The role of η is to make sure that the distance between leaf nodes in deeper level is smaller. In Figure **??**, according to Eq. (**??**), let $\theta_1 = 3$, $\eta = 2$. Thus the weight between "apple juice" and "red apple juice" could be calculated as:

$$w(apple\ juice, red\ apple\ juice) = \frac{f\left(apple\ juice\right) \cdot \frac{fr(red\ apple\ juice)}{fr(apple\ juice)}}{3^2} = \frac{1 \cdot \frac{40}{60}}{3^2} = 0.074.$$

In the same way, other weights could be calculated as shown in Figure **??**.

Let $dist(Cat_u^d, root^d)$ be the distance between Cat_u^d and the root node, so the value of $dist(Cat_u^d, root^d)$ is the sum of weights on edges from Cat_u^d to the root node. For example, in Figure **??**, $dist(coke, drink) = 0.075 + 0.333 = 0.408$.

Let Cat_u^d and Cat_v^d be two different attribute values which are the leaf nodes of the tree. Let $CAN\left(Cat_u^d, Cat_v^d\right)$ be the closest ancestor node for Cat_u^d, Cat_v^d. For example, in Figure **??**, $CAN\left(pepsi, coke\right) = carbonated$, $CAN\left(boiled\ water, orange\right) = Drink$.

Let $Cdist(Cat_u^d, Cat_v^d)$ be the distance between Cat_u^d and Cat_v^d, then we have the following formula:

$$\begin{aligned} &Cdist(Cat_u^d, Cat_v^d) = \\ &\left|dist(Cat_u^d, root^d) + dist(Cat_v^d, root^d)\right. \\ &\left.-2 \times dist(CAN\left(Cat_u^d, Cat_v^d\right), root^d)\right| \end{aligned} \tag{8}$$

For instance, in Figure **??**, the distance between "strawberry juice" and "red apple juice" is calculated as:

$$\begin{aligned} &Cdist\left(strawberry\ juice, red\ apple\ juice\right) \\ &= |dist\left(strawberry\ juice, Drink\right) + \\ &\quad dist\left(red\ apple\ juice, Drink\right) - \\ &\quad 2 \times dist\left(juice, Drink\right)| \\ &= |(0.5 + 0.067) + (0.5 + 0.1 + 0.074) - 2 \times 0.5| = 0.241 \end{aligned}$$

In the same way, we can compute the distance between the other categorical attribute values according to Eq. **??**.

2) Ordinal values: The measurement for ordinal attribute values is similar to categorical attributes, which is also based on a weighted tree structure. The distance between two attribute values is the length of the weighted path linked the two values. The computing process concludes three similar steps as used in the measurement of categorical attributes.

Let Ord^d be the d-th ordinal attribute, its attribute values have meaning of order or rank. For example, let us observe an ordinal attribute, $degree$, its attribute values conclude {Very Bad, Bad, No Comment, Good, Very Good}. Set y be the size of the domain of Ord^d. In the example, $y = 5$. According to the semantic of attribute values, the weighted tree could be constructed as shown in Figure **??** and there are five nodes. That is to say, the $root$ of the tree is "No Comment", the left branch is about the negative rank, the right branch is the positive rank. The smaller rank one has the smaller depth. In fact, the structure also can be seen as a line, which is a kind of special tree.

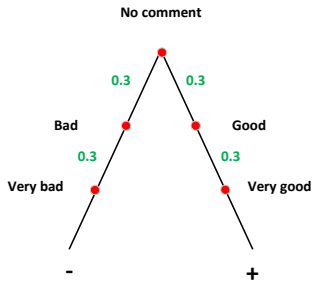

Figure 2. The tree structure of ordinal values

Let $w\left(Ord^d\right)$ be the weight function between two adjacent nodes in the tree. Inspired by Eq. **??**, the $w\left(Ord^d\right)$ could be defined as:

$$w\left(Ord^d\right) = \begin{cases} 1, & y \le \vartheta_1 \\ 1 - \zeta_1\left(y - \vartheta_1\right), & \vartheta_1 < y \le \vartheta_2 \\ 1 - \zeta_1\left(\vartheta_2 - \vartheta_1\right) - \zeta_2\left(y - \vartheta_2\right), & y > \vartheta_2 \end{cases} \tag{9}$$

where $\zeta_1, \zeta_2 \in (0, 1)$, $\vartheta_1, \vartheta_2 \in N+$.

In fact, the weight between any two adjacent nodes is same by Eq. **??**. The main reason to regard the above three situations is to harmonize the proportion among different type values. For example, in Figure **??**, if $\vartheta_1 = 3, \vartheta_2 = 10$

and $\zeta_1 = 0.35$, the weight between two adjacent nodes is 0.3.

Let Ord_u^d and Ord_v^d be two different attribute values in the tree. Let LMN be the left-most node. For example, in Figure **??**, $LMN^d = Very\ Bad$. Let $dist(Ord_u^d, LMN^d)$ be the distance between Ord_u^d and LMN^d. The value of $dist(Ord_u^d, LMN^d)$ is the sum of weights on edges from Ord_u^d to LMN^d. For example, in Figure **??**, $dist(Good, LMN^d) = dist(Good, Very\ Bad) = 0.9$.

Let $Odist(Ord_u^d, Ord_v^d)$ be the distance between Ord_u^d and Ord_v^d. It is calculated by the following formula:

$$\begin{aligned} &Odist(Ord_u^d, Ord_v^d) = \\ &|dist(Ord_u^d, LMN^d) - dist(Ord_v^d, LMN^d)|. \end{aligned} \tag{10}$$

In the example, we have the distance between $Good$ and Bad as:

$$\begin{aligned} &Odist\,(Good, Bad) = \\ &|dist(Good, Very\ Bad) - dist(Bad, Very\ Bad)| \\ &= |0.9 - 0.3| = 0.6. \end{aligned}$$

3) Numerical values: In many cases, the distance between any two categorical or ordinal attributes is relatively small to the distance between numerical values. Therefore, the difference produced by categorical and ordinal attributes will be eliminate if the numerical attribute values are too large. In order to decrease the influence, it is necessary to carry out normalization for numerical attributes.

Let Num^d be the d-th numerical attribute, and Num_i^d is the i-th attribute value. μ represents the mean of its attribute values, δ is the standard deviation. The Z-score normalization function is as follow:

$$f_{num^d}\left(Num_i^d\right) = \frac{Num_i^d - \mu}{\delta}. \tag{11}$$

Let $Ndist(Num_u^d, Num_v^d)$ be the distance between Num_u^d and Num_v^d. Then, it is computed by the following formula:

$$\begin{aligned} &Ndist(Num_u^d, Num_v^d) \\ &= \left|f_{num^d}\left(Num_u^d\right) - f_{num^d}\left(Num_v^d\right)\right|. \end{aligned} \tag{12}$$

4) Distance between objects: For any object $\mathbf{x_n}$ in a universe $\mathbf{U}$, where $\left\{x_n^1, \cdots, x_n^P\right\}$, $\left\{x_n^{P+1}, \cdots, x_n^{P+Q}\right\}$, $\left\{x_n^{P+Q+1}, \cdots, x_n^{P+Q+T}\right\}$ represent the categorical attribute values, ordinal attribute values, numerical attribute values of $\mathbf{x_n}$ respectively. P, Q and T represent the number of different attribute types.

Let $Dist\,(\mathbf{x}_i, \mathbf{x}_j)$ be the distance between objects $\mathbf{x_i}$ and $\mathbf{x_j}$. We have the measurement of distance as follows.

$$\begin{aligned} Dist\,(\mathbf{x}_i, \mathbf{x}_j) = &\sqrt{\sum_{p=1}^{P} Cdist^2\left(x_i^p, x_j^p\right)} \\ &+ \sqrt{\sum_{q=1}^{Q} Odist^2\left(x_i^{P+q}, x_j^{P+q}\right)} \\ &+ \sqrt{\sum_{t=1}^{T} Ndist^2\left(x_i^{P+Q+t}, x_j^{P+Q+t}\right)}. \end{aligned} \tag{13}$$

B. The three-way clustering algorithm

In this subsection, we will adopt an evaluation function-based three-way cluster model [**?**], which produces three regions by using an evaluation function and a pair of thresholds on the values of the evaluation function. Suppose there are a pair of thresholds (α, β) and $\alpha \geq \beta$. Although evaluations based on a total order are restrictive, the model based on two thresholds has a computational advantage. One can obtain the three regions by simply comparing the evaluation value with a pair of thresholds. Based on an evaluation function $v(\mathbf{x})$, we get the following three-way decision rules:

$$\begin{aligned} Co(C_k) &= \{\mathbf{x} \in \mathbf{U} | v(\mathbf{x}) \geq \alpha\}, \\ Fr(C_k) &= \{\mathbf{x} \in \mathbf{U} | \beta \leq v(\mathbf{x}) < \alpha\}, \\ Tr(C_k) &= \{\mathbf{x} \in \mathbf{U} | v(\mathbf{x}) < \beta\}. \end{aligned} \tag{14}$$

In fact, the evaluation function $v(\mathbf{x})$ can be a risk decision function, a similarity function, a distance function and so on. In other words, the evaluation function will be specified accordingly when an algorithm is devised.

In this paper, a three-way clustering algorithm for mixed-type data (shorted by TWD-MD) is proposed and depicted in Algorithm **??**. The basic idea of the algorithm is to find the center of K clusters first. In fact, there is a bunch of clustering approaches to determine the center. In our experiments, we adopt the outstanding density peaks clustering method [**?**]. Then, the left work is to decide the left objects where to go. Line 3 to Line 8 describe how to decide the left objects to the core region or fringe region of corresponding cluster.

In order to make decisions, we find the neighbors $X_{i-Neighbor}$ within the neighbor radius R_{th} of the object $\mathbf{x}_i$. The neighbors are find by the following formula:

$$X_{i-Neighbor} = \{\mathbf{x}_j | dist(\mathbf{x}_i, \mathbf{x}_j) \leq R_{th}\}. \tag{15}$$

Then, the object $\mathbf{x}_i$ is assigned to the core region or fringe region of the corresponding clusters according to the proportion of each cluster in the neighbor objects set $X_{i-Neighbor}$. That is, the proportion is defined as follows:

$$P(X_{i-Neighbor}|C_k) = \frac{|\{\mathbf{x}_j | \mathbf{x}_j \in X_{i-Neighbor} \wedge \mathbf{x}_j \in C_k\}|}{|X_{i-Neighbor}|}. \tag{16}$$

According to the above formula, the three-way decision rules is given as follows:

$$\begin{aligned} &if\ P(X_{i-Neighbor}|C_k) \geq \alpha, \\ &\quad the\ object\ is\ decided\ to\ Co(C_k); \\ &if\ \beta \leq P(X_{i-Neighbor}|C_k) < \alpha, \\ &\quad the\ object\ is\ decided\ to\ Fr(C_k); \\ &if\ P(X_{i-Neighbor}|C_k) < \beta, \\ &\quad the\ objecis\ decided\ to\ Tr(C_k). \end{aligned} \tag{17}$$

How to decide the threshold α and β automatically is still a unsolved problem. We can decide the thresholds by experience or through active learning method in future work.

Algorithm 1: the three-way clustering algorithm for mixed-type data

Input: K, R_{th}, α, β;
Output: $\mathbf{C} = \{\{Co(C_1), Fr(C_1)\}, \cdots, \{Co(C_K), Fr(C_K)\}\}$.

```
to compute the distance matrix between objects using Formula ??);
to obtain the K center and the initial two-way clustering result using the method in [?];
for every x_i which is not a center do
    to computer X_{i-Neighbor} according to Eq. ??);
for every C_k in the initial result do
    for every x_i do
        to computer P(X_{i-Neighbor}|C_k) by Eq. ??;
        to decide the object to the corresponding core region, fringe region or trivial region according to the three-way rules ??.
```

IV. Experiments

In this section, we validate the proposed method on some real-world datasets. The proposed algorithm is implemented in Visual Studio 2012 development environment using C++ programming language. All experiments are tested in a PC with Intel(R) Core(TM) i5-4430S CPU @ 2.70GHz, 8G RAM.

A. Three-way clustering result simulations

In order to visualize intuitively the difference between three-way clustering and the traditional two-way clustering, a synthetic two dimensions data set with 1000 points is employed in this subsection. The two-way clustering algorithm is the method in [?]. The results are depicted in Figure ?? and Figure ??, respectively.

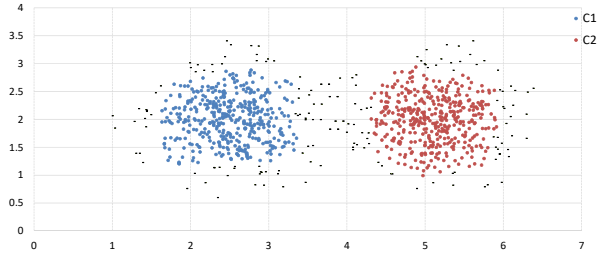

Figure 3. The result of the reference [?] on a synthetic dataset

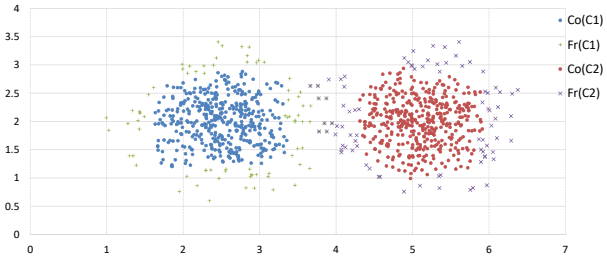

Figure 4. The three-way clustering result on a synthetic dataset

From Figure ??, we see that these points are clustered into two clusters and each point is belong to only one cluster. Figure ?? shows that compared with the traditional two-way clustering, three-way clustering algorithm represents the core region and the fringe region of every cluster correctly. Of course, the overlapping ones of two classes are also shown up. There are 80 points in the fringe region of clusters C1 and 85 points in the fringe region of clusters C2, and the two clusters have 8 overlapping points.

B. The accuracy on UCI data sets

For mixed-type data clustering, the distance measure between objects directly affects the correctness of clustering results. In turn, the accuracy of clustering results reflects the performance of measurement. Thus, we test the accuracy on the proposed method TWD-MD and the compared algorithms such as k-means mixed algorithm [?] and Paired k-means algorithm [?].

We need to note that the results of the compared algorithm are from the original references respectively. TABLE ?? describes the information of five data sets from UCI data [?]. The comparative results are shown in Figure ??.

From Figure ??, we see that the accuracy of the proposed algorithm is higher than the contrastive algorithms except for data set Iris. Let us observe the five data sets, we find that the Iris only has numerical attribute values. This shows that the proposed measurement, which is based on the weighted tree structure, has more precise measurement effect on mixed-type data.

Table I
DATASETS CHARACTERISTICS

Data sets	Size	Attribute Number			Clusters
		Cat	Ord	Num	
Iris	150	0	0	4	3
Teaching Assistant Evaluation	151	4	0	1	3
Congressional Voting Records	435	16	0	0	2
Credit Approval	690	9	0	6	2
Adult	48842	7	1	6	2

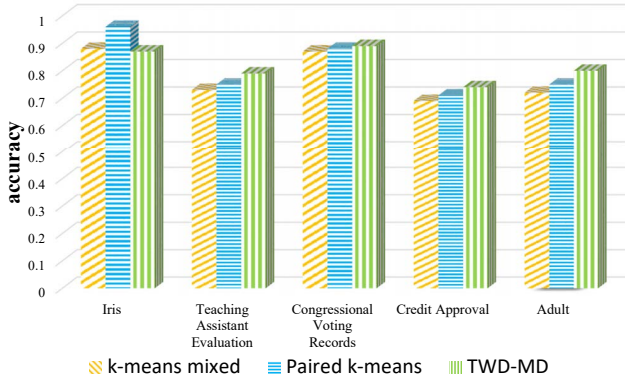

Figure 5. The results of different algorithms on Accuracy

C. The ARI on UCI data sets

For mixed-type data clustering, the Average adjusted Rand Index (ARI) is another important indicator to estimate the clustering result. The higher the value of ARI is, the more accurate the clustering result is.

In this subsection, we compare the proposed algorithm TWD-MD with CoupledMC algorithm [**?**] and EGMCM algorithm [**?**]. We also adopt the results of the compared algorithm from the original references. TABLE **??** gives the information of the five data sets. The values of ARI of the proposed algorithm and contrastive algorithms are shown in TABLE **??**.

Table II
DATASETS CHARACTERISTICS

Data sets	Size	Attribute Number			Clusters
		Cat	Ord	Num	
Heart Disease (Switzerland)	303	5	3	5	2
Heart Disease (VA)	303	5	3	5	5
Climate Model Simulation Crashes	540	0	0	18	2
QSAR BioCdegradation	1055	8	8	25	2
Contraceptive Method Choice	1473	4	4	1	3

Table III
THE RESULTS OF DIFFERENT ALGORITHMS ON ARI

Data sets	CoupledMC	EGMCM	TWD-MD
Heart Disease (Switzerland)	0.18	**0.26**	0.24
Heart Disease (VA)	0.02	0.12	**0.15**
Climate Model Simulation Crashes	0.01	0.11	**0.2**
QSAR BioCdegradation	0	0.23	**0.48**
Contraceptive Method Choice	0.04	0.01	**0.19**

From TABLE **??**, we see that the ARI of the proposed algorithm is the best one in most cases. It is shown that the proposed method is good at clustering mixed-type data.

V. CONCLUSIONS

The data is characterized by numerical attributes, categorical attributes or ordinal attributes. The main objective of this paper is to propose a new way for clustering the mixed-type data. Thus, we proposed different measurements of distance for different types of attribute values by considering the semantic of attributes, the number of attribute values and the occurrence frequency of attribute values. The measurements borrow the representation of weighted tree structures.

Besides, we constructed the clustering result based on the representation of three-way cluster, which is suitable for dealing with uncertainty because it shows intuitively which objects are fringe to the cluster. That is, a cluster is presented by a pair of sets instead of a single set as used by many existing methods. Three regions just reflect the relationships between an object and a cluster, namely, an object definitely in a cluster, an object definitely not in a cluster, an object might be in a cluster or not.

Then, a three-way clustering algorithm for mixed-type data was proposed. The algorithm first finds the center of K clusters by using a method such as the outstanding density peaks clustering method [**?**]. Then, we build the three-way decision rules based on the proposed measurement of distance to obtain the three-way clustering result. The experimental results show that the proposed distance measure is reasonable and effective for mixed-type data, and the proposed algorithm is in a better performance at the accuracy and the average adjusted rand index than the compared algorithms in most cases.

We also find that the proposed algorithm is not always the best one though it is doing well for mixed-type data sets. The thresholds used in the experiments are set by experience, how to decide them automatically is one of the further research work.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 61379114 , 61533020 & 2016YFB1000905.

REFERENCES

[1] H. Frigui, O. Bchir, and N. Baili, "An overview of unsupervised and semi-supervised fuzzy kernel clustering," *International Journal of Fuzzy Logic and Intelligent Systems*, vol. 13, no. 4, pp. 254–268, 2013.

[2] Z. Kang, C. Peng, and Q. Cheng, "Twin learning for similarity and clustering: A unified kernel approach," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17), AAAI Press*, 2017.

[3] Z. X. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data mining and knowledge discovery*, vol. 2, no. 3, pp. 283–304, 1998.

[4] C. Ning, C. An, and Z. Long Xiang, "Fuzzy k-prototypes algorithm for clustering mixed numeric and categorical valued data," *Journal of software*, vol. 12, no. 8, pp. 1107–1119, 2001.

[5] J. C. Ji, W. Pang, C. G. Zhou, X. Han, and Z. Wang, "A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data," *Knowledge-Based Systems*, vol. 30, pp. 129–135, 2012.

[6] J. C. Ji, T. Bai, C. G. Zhou, C. Ma, and Z. Wang, "An improved k-prototypes clustering algorithm for mixed numeric and categorical data," *Neurocomputing*, vol. 120, pp. 590–596, 2013.

[7] A. Pathak and N. R. Pal, "Clustering of mixed data by integrating fuzzy, probabilistic, and collaborative clustering framework," *International Journal of Fuzzy Systems*, vol. 18, no. 3, pp. 339–348, 2016.

[8] J. Y. Chen and H. H. He, "A fast density-based data stream clustering algorithm with cluster centers self-determined for mixed data," *Information Sciences*, vol. 345, pp. 271–293, 2016.

[9] D. Lam, M. Z. Wei, and D. Wunsch, "Clustering data of mixed categorical and numerical type with unsupervised feature learning," *IEEE Access*, vol. 3, pp. 1605–1613, 2015.

[10] N. Aziz, M. T. Ismail, S. Ahmad, and R. A. Rahman, "Robust influence angle for clustering mixed data sets," in *AIP Conference Proceedings*, vol. 1605, no. 1. AIP, 2014, pp. 840–843.

[11] C. C. Hsu, S. H. Lin, and W. S. Tai, "Apply extended self-organizing map to cluster and classify mixed-type data," *Neurocomputing*, vol. 74, no. 18, pp. 3832–3842, 2011.

[12] W. S. Tai and C. C. Hsu, "Growing self-organizing map with cross insert for mixed-type data clustering," *Applied Soft Computing*, vol. 12, no. 9, pp. 2856–2866, 2012.

[13] S. Boriah, V. Chandola, and V. Kumar, "Similarity measures for categorical data: A comparative evaluation," in *Proceedings of the 2008 SIAM International Conference on Data Mining*. SIAM, 2008, pp. 243–254.

[14] D. W. Goodall, "A new similarity index based on probability," *Biometrics*, pp. 882–907, 1966.

[15] F. Noorbehbahani, S. R. Mousavi, and A. Mirzaei, "An incremental mixed data clustering method using a new distance measure," *Soft Computing*, vol. 19, no. 3, pp. 731–743, 2015.

[16] Y. Zhang, F. Zhang, B. Q. Zhu, Z. M. Xiang, and L. Tang, "A new method of color pattern recognition based on fuzzy clustering," in *AER-Advances in Engineering Research*, vol. 64, 2016, pp. 135–138.

[17] H. Yu and Y. Wang, "Three-way decisions method for overlapping clustering," in *Proceedings of the 8th International Conference on Rough Sets and Current Trends in Computing*, 2012, pp. 277–286.

[18] H. Yu, P. Jiao, Y. Y. Yao, and G. Y. Wang, "Detecting and refining overlapping regions in complex networks with three-way decisions," *Information Sciences*, vol. 373, pp. 21–41, 2016.

[19] H. Yu, C. Zhang, and G. Y. Wang, "A tree-based incremental overlapping clustering method using the three-way decision theory," *Knowledge-Based Systems*, vol. 91, pp. 189–203, 2016.

[20] Y. Y. Yao, "Three-way decision: an interpretation of rules in rough set theory," in *International Conference on Rough Sets and Knowledge Technology*. Springer, 2009, pp. 642–649.

[21] ——, "The superiority of three-way decisions in probabilistic rough set models," *Information Sciences*, vol. 181, no. 6, pp. 1080–1096, 2011.

[22] H. Yu, G. Y. Wang, and etc., *Methods and Practices of Three-Way Decisions for Complex Problem Solving*. Springer International Publishing, 2015.

[23] C. C. Hsu, K. M. Wang, and S. H. Wang, "Gvisom for multivariate mixed data projection and structure visualization," in *Neural Networks, 2006. IJCNN'06. International Joint Conference on*. IEEE, 2006, pp. 3300–3305.

[24] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.

[25] A. Ahmad and L. Dey, "A k-mean clustering algorithm for mixed numeric and categorical data," *Data & Knowledge Engineering*, vol. 63, no. 2, pp. 503–527, 2007.

[26] H. Haripriya, S. Amrutha, R. Veena, and P. Nedungadi, "Integrating apriori with paired k-means for cluster fixed mixed data," in *Proceedings of the Third International Symposium on Women in Computing and Informatics*. ACM, 2015, pp. 10–16.

[27] https://archive.ics.uci.edu/ml/datasets.html.

[28] C. Wang, C. H. Chi, W. Zhou, and R. K. Wong, "Coupled interdependent attribute analysis on mixed data." in *AAAI*, 2015, pp. 1861–1867.

[29] S. B. Vaibhav Rajan, "Dependency clustering of mixed data with gaussian mixture copulas." IJCAI, 2016, pp. 1967–1973.