01 Sep 2015

Clustering Data of Mixed Categorical and Numerical Type with Unsupervised Feature Learning

Dao Lam

Mingzhen Wei
Missouri University of Science and Technology, weim@mst.edu

Donald C. Wunsch
Missouri University of Science and Technology, dwunsch@mst.edu

Recommended Citation

D. Lam et al., "Clustering Data of Mixed Categorical and Numerical Type with Unsupervised Feature Learning," *IEEE Access*, vol. 3, pp. 1605-1613, Institute of Electrical and Electronics Engineers (IEEE), Sep 2015.
The definitive version is available at https://doi.org/10.1109/ACCESS.2015.2477216

Received August 1, 2015, accepted August 28, 2015, date of publication September 7, 2015, date of current version September 25, 2015.

Digital Object Identifier 10.1109/ACCESS.2015.2477216

Clustering Data of Mixed Categorical and Numerical Type With Unsupervised Feature Learning

DAO LAM[1], (Member, IEEE), MINGZHEN WEI[2], AND DONALD WUNSCH[1], (Fellow, IEEE)

[1]Applied Computational Intelligence Laboratory, Department of Electrical and Computer Engineering, Missouri University of Science and Technology, Rolla, MO 65401, USA

[2]Department of Geological Science and Engineering, Missouri University of Science and Technology, Rolla, MO 65409, USA

Corresponding author: D. Lam (dlmg4@mst.edu)

This work was supported in part by the Missouri Science and Technology Intelligent Systems Center and in part by Mary K. Finley Missouri Endowment.

ABSTRACT Mixed-type categorical and numerical data are a challenge in many applications. This general area of mixed-type data is among the frontier areas, where computational intelligence approaches are often brittle compared with the capabilities of living creatures. In this paper, unsupervised feature learning (UFL) is applied to the mixed-type data to achieve a sparse representation, which makes it easier for clustering algorithms to separate the data. Unlike other UFL methods that work with homogeneous data, such as image and video data, the presented UFL works with the mixed-type data using fuzzy adaptive resonance theory (ART). UFL with fuzzy ART (UFLA) obtains a better clustering result by removing the differences in treating categorical and numeric features. The advantages of doing this are demonstrated with several real-world data sets with ground truth, including heart disease, teaching assistant evaluation, and credit approval. The approach is also demonstrated on noisy, mixed-type petroleum industry data. UFLA is compared with several alternative methods. To the best of our knowledge, this is the first time UFL has been extended to accomplish the fusion of mixed data types.

INDEX TERMS Clustering, unsupervised feature learning, mixed-type data, fuzzy ART.

I. INTRODUCTION

Our work addresses the problem of mixed-type categorical and numerical data in clustering. The goal is building a framework that automatically handles the differences in numerical and categorical features in a dataset and groups them into similar clusters.

Clustering is the problem of grouping unlabeled data items into classes based on the similarity of the items [1]. Many clustering algorithms, such as K-means and spectral clustering [2], assume that features have numeric values. However, in practice, mixed, erroneous, and missing data can result from i) errors or mistakes caused by the equipment or humans, or ii) the data attributes, which can be either numerical or categorical. Combinations of these issues can cause data to be mixed-type, multivalued, or missing. This paper presents a mechanism for overcoming these problems.

Many other algorithms, such as those discussed in [3] and [4], are designed only for categorical data. Methods for handling mixtures of attributes were researched and discussed in [5] and [6], but these approaches are just the beginning of what is needed. They typically group the attributes as categorical or numerical and treat them separately until finally combining them using a distance function. Li and Biswas [5] demonstrated a similarity measure based on biometric classification, in which greater weight is assigned to features that are uncommon in the population. Based on that distance, they proposed the hierarchical agglomerative algorithm named Similarity-based Agglomerative Clustering (SBAC). However their quadratic computational cost is expensive. Many researchers, such as [7] and [8], have extended the K-means algorithm to work with data containing both numerical and categorical features but still treat them separately. In [9], numerical and categorical features were treated separately during the clustering process, and the results then were combined to obtain a better partition using the ensemble learning approach. Hsu and Wang [10] and Hsu and Chen [11] proposed a variance and entropy clustering for mixed data but it requires domain expertise to build

the distance hierarchy for categorical attributes. In this paper, a new approach based on UFL allows a seamless combination of both categorical and numerical features. To our knowledge, this the first time that a method has been developed to extend UFL capabilities to mixed-type data. We accomplish this by combining it with the data fusion capabilities of Fuzzy ART. This is also the first time we are aware of ART being with the UFL technique.

In addition to being mixed, categorical data can have multiple values, and numerical data can have a range of values. Very little research has been conducted regarding this problem. In [12], the investigators addressed the problem of multi-value data in database clustering by building the similarity as the combination of qualitative and quantitative features. Other researchers [13] have used the Hausdorff distance to compute the distance between interval features, followed by a dynamic clustering algorithm to cluster the dataset. Such approaches are limited to either discrete or continuous features.

One of the main challenges of clustering is to determine the true number of clusters in a dataset. Several algorithms, such as K-means and spectral clustering [2], assume that this number must be known a priori. Other methods, such as fuzzy ART clustering [14], do not require this information, but the ideal number of clusters often is determined during the cluster validation process, and numerous studies have contributed to the solution [1], [15].

UFL has been widely used in computer vision [16], [17]. Besides the advantage of removing the labor of designing application specific features, results confirm that UFL shows higher performance than traditional approaches such as discussed in [18] and [19]. UFL uses one of the unsupervised learning algorithms, such as K-means or auto encoding, to learn the features but those clustering methods often require numerical data. To the best of our knowledge, there is no previous research on applying UFL to mixed-type feature data.

This paper uses ART as the unsupervised learning algorithm to learn the features from the data itself. ART has many attractive characteristics. It scales very well for large scale datasets because of its low computational requirement which is *O(NlogN)* or it can be further reduced to *O(N)* [14] when in a one-pass learning mode. The other reason why ART is chosen as UFL for mixed-type data is its ability in data fusion [20], [21] by mapping features from multimodal data simultaneously. Moreover, ART can dynamically and adaptively generate a prototype, which is used in feature encoding, without the requirement of specifying the number of clusters.

In this paper, a novel approach to handle mixed-type data clustering is presented by using UFL. The contributions of this paper are:

1. It presents a UFL approach using Fuzzy ART. Unlike other unsupervised learning methods like K-means or auto encoder [22], where one needs large amounts of data, the approach presented here works for both large and small volumes of data, which can be relevant when some subspaces of the data are represented with many samples and other subspaces are represented by relatively few samples.

2. UFLA can solve the problem of mixed-type data. By learning the higher and sparse feature representation, the distinction between categorical and numerical in the original data becomes less of an obstacle.

The approach is tested on several datasets with mixed features: heart disease, teaching assistant evaluation and credit assignment on UCI repositories [23]. We also test one dataset with no ground truth from the petroleum industry [24].

The rest of the paper is organized as follows: Sec. II includes a review of UFL and Fuzzy ART, and Sec. III presents our novel approach to solve the problem of mixed, erroneous, and missing features. Sec. IV describes our experiments with real data from the UCI machine learning repository and the petroleum industry. Finally, conclusions and some future research directions are discussed in Sec. V.

II. MOTIVATION AND BACKGROUND

A. NOVELTY AND MOTIVATION

Although UFL can be widely applied in many areas, its approach has never been investigated as being applicable to mixed-type data representing real life datasets.

UFL is known to successfully represent the object in another sparse representation [25]. UFL can discover hidden features in data and represents them in sparse domains, which are more suitable for machine learning tasks than the original data.

The motivation lies in applying the UFL to mixed-type data to reduce the distinction between the numerical and categorical. Most UFL [22], [26] has traditionally served as a preprocessing method for supervised learning problems. This leaves open the opportunity to apply it to a purely unsupervised problem. Thus, this paper investigates UFL's contribution to the hard but important problem of clustering when dealing with numeric and categorical data.

This motivation leads to the question of how to apply UFL to mixed-type dataset. The novelty of this method is the use of Fuzzy ART, one method of unsupervised learning, as the method of building a feature encoder.

The difference between categorical and numerical features makes several traditional clustering methods fail since they often work with numbers only. Many approaches try to treat them separately and then combine them in a later step [9] but they are still treated differently, and it is unclear whether the end results are satisfactory. An ideal strategy would be to fuse categorical and numerical features before actual clustering. UFL is very successful in sparse representation of the objects in a different space [27]. Unfortunately, that type of UFL requires a large amount of data to build the encoder. Those large datasets are affordable when working with images, and video, since their natural features are large dimension and require convolutional processing. For those real datasets as shown in the UCI datasets used in this

paper's experiments, each sample is represented by less than a few dozen features, and the convolutional operation is unworkable.

This leads to the introduction of Fuzzy ART into the process of UFL. This paper investigates a new method of UFL using Fuzzy ART. ART is known for its fast performance in unsupervised learning [14]. The other main advantage of Fuzzy ART is it does not need to specify the number of clusters. The other motivation for using ART to resolve the distinction between numerical and categorical features is its capability of data fusion as demonstrated in [21].

The next two sections review the related background that is necessary for the new approach.

B. UFL

Traditional classifiers require a human operator to high-level features, such as the scale-invariant feature transform (SIFT) and histogram of oriented gradients (HOG) [28], [29]. Those approaches are difficult or time consuming to apply to other kinds of data. To tackle the disadvantage of hand-crafted features, several methods of UFL have been researched, such as sparse coding, deep belief nets, auto encoder, and independent subspace analysis [16], [26].

In machine learning, the amount of data is often more important than the choice of algorithm. This is especially true in UFL where simple learning algorithms outperform several handcrafted, carefully designed methods [30]. Recently many researchers have started using UFL in computer vision, e.g. [17] used sparse coding, and [22] used one-layer UFL for classification of text and image.

In tasks such as image classification and object recognition, UFL can be a more attractive approach than those relying on manually-designed features [22], [31]. UFL has also proven to be helpful in greedy layer-wise pre-training of deep architectures [32]–[34].

However, a major drawback of many UFL systems is their complexity where parameters like learning rate, momentum, and weight decay must be tuned and network architecture parameters must be cross-validated. This paper investigates a new method of UFL using a simple, fast but effective training by using Fuzzy ART. While most other UFL algorithms focus on applying to classification, this work serves to reduce the gap in numerical and categorical features and work under the clustering domain, an unsupervised learning problem. ART has shown its ability in data fusion by mapping multimodal features in an incremental manner [20].

More over, UFL often leads to sparse feature representation, as demonstrated in [25]. Sparse representation often has several advantages such as robustness to noise. For clustering, sparse representation is probably easier to separate in higher dimensional space.

An unsupervised learning task often consists of four broad steps: 1) feature extraction 2) feature encoder building 3) feature mapping and 4) feature pooling. Our approach to feature learning removes the feature extraction and feature pooling because they are not relevant to the mixed-type data. We only keep feature encoder building with Fuzzy ART clustering and feature mapping with a soft threshold function where the weight below a certain threshold is set to 0 resulting in sparse representation features.

C. FUZZY ART

ART has been applied successfully to many machine learning applications [35]. It has the advantage of fast and stable learning. It is an online learning algorithm so it can be very scalable for large scale datasets. ART also has noise immunity in document clustering [36].

The other advantage of Fuzzy ART is that ART can be used for data fusion by extending ART from a single input field to multiple ones [20] as Fusion ART. Fusion ART provides a general mechanism for multi-channel features mapping. Meng *et al.* [21] show that Fusion ART works successfully in integrating visual and textual features for image text co-clustering.

Let $\mathbf{X} = \{\mathbf{x}_1, \ldots \mathbf{x}_N\}$ be a set of N samples in the given dataset, where $\mathbf{x}_i = [x_{1,i}, \ldots x_{d,i}]^T$ is a sample belonging to d-dimensional space R^d.

The basic module of this UFL for mixed-type clustering is Fuzzy ART [14]. Fuzzy ART consists of two layers of neurons: the input layer F1 and the clustering representation layer F2. Unlike ART 1, where there are bottom up and top down weight vectors, Fuzzy ART has only one weight vector $\mathbf{w}_j$ for each category j, which is initialized to $w_{j,1} = w_{j,2} = \cdots = 1$ when the category is uncommitted.

Before the samples can be input to ART, it has to be normalized to [0, 1] and enhanced with complement coding to avoid category proliferation problems. The clusters are formed in layer F2. When an input $\mathbf{x}$ is presented to layer F1, the committed neurons and one uncommitted neuron compete in a winner take all manner to select the one with maximum activation according to the formula below:

$$Tj = \frac{|\mathbf{x} \wedge \mathbf{w}_j|}{\alpha + |\mathbf{w}_j|}, \tag{1}$$

where α is a small real number to break the tie and the fuzzy AND operator $\wedge$ is defined by

$$(\mathbf{x} \wedge \mathbf{y})_i = min(x_i, y_i), \tag{2}$$

and where the norm |.| is defined by

$$|\mathbf{x}| = \Sigma_{i=1}^{d} |x_i|. \tag{3}$$

The winning neuron J, which is $argmax_j T_j$ becomes activated. If neuron J passes the vigilance ρ criterion which is:

$$\rho < \frac{|\mathbf{x} \wedge \mathbf{w}_J|}{|\mathbf{x}|}, \tag{4}$$

then the weight adaption occurs:

$$\mathbf{w}_J(new) = \gamma(\mathbf{x} \wedge \mathbf{w}_J(old) + (1 - \gamma)\mathbf{w}_J(old), \tag{5}$$

where γ is the learning rate parameter.

On the other hand if the vigilance criterion is not met, the current winning neuron is disabled, and the next winning

neuron is chosen. If the uncommitted neuron is chosen, a new uncommitted neuron is created for future learning.

The advantage of UFL using Fuzzy ART is the dynamics, which can create many prototypes used for learning feature by just increasing the value of the vigilance threshold ρ.

III. UFLA CLUSTERING WITH MIXED, ERRONEOUS, MISSING FEATURE DATA

Because $\mathbf{x}_i$ has both categorical and numerical features, it can be represented as $\mathbf{x}_i = [xc_{1,i}, xc_{2,i}, \ldots xc_{r,i}, xn_{1,i}, xn_{2,i}, \ldots, xn_{s,i}]$, where the first part, $[xc_{1,i}, xc_{2,i}, \ldots xc_{r,i}]$, is categorical, $[xn_{1,i}, xn_{2,i}, \ldots, xn_{s,i}]$ is numerical, r and s are the number of categorical and numerical features, respectively, and $r + s = d$. In other words, the dataset X has $fc_1, fc_2, \ldots, fc_r$ as categorical features and $fn_1, fn_2, \ldots, fn_s$ as numerical features.

The proposed methodology consists of the five steps described in Algorithm 1.

Algorithm 1 UFLA Clustering

1. Perform data preprocessing to clean up missing, interval, and multi-value data. Perform binary feature mapping on categorical data and normalize numeric data to [0 1].
2. Perform fuzzy ART clustering to obtain a certain number of clusters. Consider the weights from each cluster of the ART as centroids.
3. For each sample $\mathbf{x}$ compute $f(\mathbf{z}) = min(0, mean(\mathbf{z}) - \mathbf{z})$ where z is the distance from $\mathbf{x}$ to centroids.
4. Treating f(z) as an unsupervised learning feature, use VAT or clustering validation to determine number of clusters k.
5. Cluster the new dataset into k clusters via K-means to obtain the final partition.

A. CATEGORICAL FEATURES AND NUMERICAL FEATURES PREPROCESSING

Consider a categorical feature fc_u $(u = 1, \ldots r)$ that has a domain of l values $\{d_1, d_2, \ldots d_l\}$. In the binary vector $[b_1, b_2, \ldots b_l]$, each b_v corresponds to each domain value d_v. A binary feature transform of categorical feature value d_v is the assignment of the categorical value of each sample to a binary vector of d elements $[b_1, b_2, \ldots b_l]$, where all of the entries are 0, except b_v.

Binary feature transforms are used to handle multi-value categorical features by setting the corresponding entries in the binary vector. Furthermore, missing values can be resolved by setting all of the binary entries to 1.

One form of uncertainty in this feature occurs when data are specified by a range of values, instead of one scalar. To correct this problem, if a numerical feature fn_u has interval data $[a, b]$, it is represented by two numeric features, $fn_{u,1} = a$ and $fn_{u,2} = b$.

Missing values of numeric features are replaced by the average of the observed value or by the k-nearest neighbors.

B. UFLA FRAMEWORK

For mixed-type feature data, the main challenge is to deal with both discrete and continuous variables at the same time in computing the distance between two samples since all the traditional clustering methods treat the features as numerical only. The UFL is used to remove the gap between the discrete and continuous property.

To overcome this hurdle, ART clustering is used since ART can work for mixed-type data (after pre-processing), we leverage the advantage of ART clustering as the unsupervised learning algorithm. Figure 1 depicts this process of UFL.

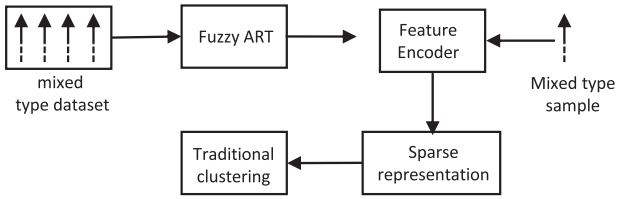


FIGURE 1. UFLA framework. The whole dataset are first clustered by Fuzzy ART to produce the prototypes of the dataset. Those prototypes were used as feature encoder to encode individual mixed-type data sample to sparse representation domain. After the mapping, the dataset can be clustered by any traditional clustering algorithm.

Moreover, ART is template based learning. The architecture summarizes the data via the examples it has seen, which makes the clusters formed represent the data structure at a specific vigilance threshold.

The algorithm sets the vigilance of the ART module to a moderate high threshold so that numerous representative clusters can be formed. N samples are then fed into the Fuzzy ART module to learn the structure of ART.

After the Fuzzy ART learning, K_F representative clusters are created and each weight $\mathbf{w}_j$ $j = 1 \ldots K_F$ connected from one cluster to the ART input is considered as the new representation of the mixed-type data. This new representation removes the gap between the numeric and categorical features.

C. UNSUPERVISED FEATURE CONSTRUCTION

To construct the unsupervised learning feature representation of a sample $\mathbf{x}$, the distance from $\mathbf{x}$ to $\mathbf{w}_j$ $j = 1 \ldots K_F$ is used to generate the UFL feature. In particular, the following feature computation is used in this paper

$$z_j = ||\mathbf{x} - \mathbf{w}_j||_2^2 \tag{6}$$

$$f_j = min(0, mean(z_1, z_2, \ldots, z_{K_F}) - z_j) \tag{7}$$

where $f = [f_1, f_2, \ldots, f_{K_F}]$ is the UFL feature representation of $\mathbf{x}$ and will be used for clustering.

D. NUMBER OF CLUSTERS

Before the K-means step in Algorithm 1 can be applied, the value of K has to be determined. Estimating the true value of K is a challenge for clustering analysis. To estimate the number of clusters we use a technique called visual assessment of tendency (VAT) [15], [37].

The VAT algorithm works by reordering the distance matrix. Each pixel intensity of the gray scale VAT image represents the dissimilarity between two samples. A black pixel means two samples are close and a white pixel means two samples are far from each other. Each object is identical to itself so the dissimilarity is 0, which is represented by a black pixel along the diagonal that has 0 intensity. The distance matrix is scaled so that the furthest distance corresponds to the white pixel with an intensity of 1. VAT uses a minimum spanning tree algorithm to organize the distance matrix so that the VAT image concentrates the dark block along the diagonal. Those dark blocks represent clusters of objects that are close to each other and the white parts that are off the diagonal represent the distances between samples in the same clusters to samples outside the cluster. VAT, therefore, can show the number of clusters along the diagonal of the VAT image.

E. CLUSTERING ANALYSIS AND EVALUATION

A critical step after clustering is analysis. This integrated methodology includes important perspectives from which to look at the clustering results.

From a statistical perspective, the number of each type of categorical feature in each cluster are counted to see which features were dominant. For numerical features, the min, max, average and standard deviation are computed. Good clustering will yield small standard deviations for each cluster, as well as averages that vary greatly from one another.

Two criteria are used to evaluate the performance of clustering. From the classification point of view, the accuracy of grouping the samples that belong to the ground truth class is computed. The resulting clusters can be classified based on the dominant number of true labels in each cluster. The average accuracy of clustering is then defined by:

$$Acc = \frac{\Sigma_i \frac{corr_i}{N_i}}{C}, \tag{8}$$

where $corr_i$, N_i are the number of correct labels and the number of objects in cluster C_i, respectively; C is the number of clusters in the dataset.

One of the most popular external clustering validation indices is the Rand index, which is defined below.

Assuming that P is the ground truth partition of dataset X with N data objects, which is also independent from a clustering structure C resulting from the use of the UFL Fuzzy ART algorithm, for a pair of data objects $\mathbf{x}_i$ and $\mathbf{x}_j$, we will have four different cases based on how $\mathbf{x}_i$ and $\mathbf{x}_j$ are placed in C and P

Case 1: $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same clusters of C and the same category of P.

Case 2: $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same clusters of C but different categories of P.

Case 3: $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to different clusters of C but the same category of P.

Case 4: $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to different clusters of C and different categories of P.

Correspondingly, the number of pairs of samples for the four cases are denoted as a, b, c, and d, respectively. Because the total number of pairs of samples is N(N–1)/2, denoted as L, we have a + b + c + d = L. The Rand index can then be defined as follows, with larger values indicating more similarity between C and P:

$$Rand = \frac{a+d}{L}. \tag{9}$$

IV. EXPERIMENT AND DISCUSSION

This section demonstrates that the methodology can perform well on several real datasets. The approach is first verified with several datasets with known ground truth: heart disease, teaching assistant, and credit assignment datasets from the UCI repository [23]. The method is also applied to a dataset without ground truth collected from the Enhanced Oil Recovery Project Survey by Oil & Gas Journal [24] to group enhanced oil recovery projects. These clustering results can help petroleum experts to better understand the data they have collected throughout years of oil production. These datasets were chosen for their challenging features. Although they are not large, they are sufficient to assess the performance of UFLA and its scalability.

A. DATASETS

1) DATASET WITH GROUND TRUTH

StatLog Heart disease dataset [23]: This UCI dataset from Cleveland Clinic has both categorical and numeric features. It has six real value features, one ordered feature (the slope of the peak exercise ST segment), and three binary features (gender, fasting blood sugar > 120 mg/dl, exercise induced angina) which can all be considered as numeric features. The rest are categorical features (resting electrocardiographic results, chest pain type, thal). Totally, it has 303 records with no missing values, 139 have heart disease and 169 do not have.

Teaching assistant evaluation dataset [23]: The dataset consists of the evaluations of teaching performance of three regular semesters and two summer semester with 151 teaching assistants at the Statistics Department of the University of Wisconsin. The scores are grouped into three groups of low, medium and high. The attributes are i) whether the TA is a native speaker ii) course instructor (25 categories) iii) course (26 categories) iv) summer or regular v) class size. This dataset is challenging since there are a lot of values for two categorical features course instructor and course.

Credit approval dataset [23]: The dataset contains 690 samples having six numeric and nine categorical features. The samples are divided into two groups, 307 approved and 383 rejected. Thirty-seven samples have missing values on seven features. This dataset is well suited for the study because it has both mixed data and missing values.

2) DATASET WITHOUT GROUND TRUTH

The petroleum data is collected from the the Oil & Gas Journal [24], biannually published for worldwide enhanced

oil recovery projects. The data in the survey were entered manually following the designed data structure. Therefore, there are several data quality problems, such as missing values, inconsistent data, erroneous data, and typos. The survey data recorded the reservoir and petroleum fluid condition, and project start year and project evaluation until the report year. Based on research on enhanced oil recovery (EOR) screening, the domain expert selected a few significant attributes for analysis, which are listed in Table 1. Among the numeric attributes, permeability, depth and viscosity had such a large dynamic range that they are represented in log scale. The main purpose of the clustering was to group data collected from several enhanced oil recovery projects. The original dataset contained a total of 460 projects.

TABLE 1. Attributes in the petroleum dataset.

Attribute	Properties
Formation Type	categorical, multi-value, missing
Porosity (%)	numerical, range value, missing
Permeability	numerical, range value, missing, log scale
Depth (ft)	numerical, range value, missing, log scale
Gravity (°API)	numerical, range value
Viscosity (cp)	numerical, range value, log scale
Temperature (°F)	numerical, range value
Residual Oil	numerical, range value

B. RESULTS AND DISCUSSION OF UCI DATASETS

In all experiments with UFL, the parameter α was fixed at 0.001 because α does not have significant influence on generating clustering nodes in the F2 layer. The learning rate γ is set at 0.9 for moderately fast learning. If γ is set at 1 (fast learning), the number of clusters in Fuzzy ART is often small and the unsupervised learning features are not meaningful. On the other hand if γ is small (slow learning), the unsupervised features are stable but the performance is slow. The main parameter to adjust in UFLA is the vigilance parameter ρ. Unlike the Fuzzy ART clustering problem where the vigilance has to be fine tuned to get the number of clusters equal to the true number of cluster in the dataset, in UFL, the vigilance adjusts roughly so that the number of clusters generated approximates the desired number of features that should be enough for the K-means steps. The vigilance values reported in Table 2 are representative only, other values of vigilance might result in the same performance. The K-means clustering at step 5 in Algorithm 1 is repeated ten times and the one with the smallest objective function is used for final clustering.

TABLE 2. Vigilance parameter and number of UFL features in heart disease, teaching assistant and credit approval datasets.

	Vigilance	Number of UFL features
Heart disease	.25	29
Teaching assistant	.60	8
Credit assignment	.60	73

For comparison, several algorithms that can handle mixed-type features are applied to the above datasets. These include K-prototypes [6], K-medoids [38]. Furthermore, since the proposed approach is based on Fuzzy ART, Fuzzy ART [14] clustering is also compared to demonstrate how the performance is improved.

Table 3 shows clearly the superior performance of UFLA clustering compared to the rest of algorithms. For the credit dataset, it has an accuracy of 86% well above the next highest accuracy, which is 79%. The teaching assistant evaluation dataset is a challenging dataset since there are many categorical values but the UFLA is still better than the other algorithms.

To motivate UFLA as an approach for dealing with mixed-type data, it is interesting to compare UFLA clustering with Fuzzy ART itself. All three datasets clearly show the effectiveness of the approach since the unsupervised features have a better representation of the mixed-type data than the original data. The reason for this higher performance is the UFL features have removed the gap between the categorical and numerical features leading to a better clustering even when using with K-means in a later stage. In the original form of data, although after preprocessing the data is in numerical form, the transformation in many cases make it hard to interpret the distance between the two samples [5].

The heart disease dataset has 303 samples and some missing values [23], which make it a good fit for this approach. Before running Algorithm 1, nominal missing values are replaced by the mode of the observed values and numerical missing values are replaced by the mean of the observed values. The FuzzyART clustering is run with vigilance 0.25, to obtain 29 unsupervised learning features.

The heart disease dataset has been used as benchmark in several mixed-type data clustering such as COBWEB/3 [39], ECOBWEB [40] and SBAC [5]. Table 4 gives the clustering partitions with confusion matrix and average accuracy of UFLA and COBWEB/3, ECOBWEB, SBAC. The UFL approach has the best performance among the algorithms.

C. RESULTS AND DISCUSSION OF PETROLEUM DATASET

1) CLUSTERING PRE-PROCESSING

For numerical features, missing values are populated by the average of the non-missing values of the respective features. Features with interval values were split into two features, one for the lower bound and one for the upper bound. To deal with noisy data, we use whisker plots to define the noisy values and treat them as missing values.

For categorical features, missing values are populated by the mode of the observed value of the category.

Binary transform is then applied to formation type features. Permeability, depth and viscosity features are transformed to log scale. All of features are then scaled into range [0, 1] for fuzzy ART clustering.

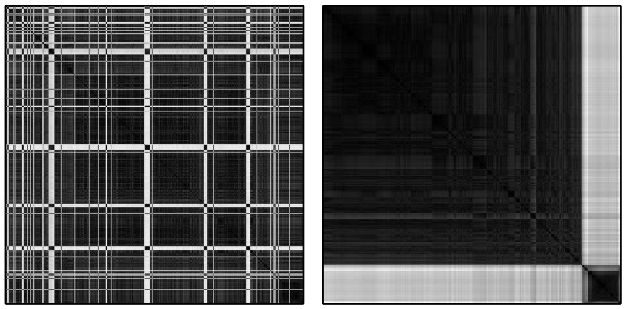
Then we applied Algorithm 1 to the pre-processed dataset. The vigilance value was set at 0.8. There are 29 clusters formed, corresponding to 29 unsupervised features learned.

TABLE 3. Performance comparison for mixed-type data clustering of K-prototype, K-medoids, Fuzzy ART and UFL Fuzzy ART.

	UFL Fuzzy ART		K-prototype		K-medoids		Fuzzy ART	
	Acc	Rand	Acc	Rand	Acc	Rand	Acc	Rand
Heart disease	81.5	69.7	80.0	63.8	76.5	61.1	46.6	50.4
Teaching assistant	52.2	59.1	40.2	55.3	46.1	53.1	44.2	50.9
Credit assignment	86.0	75.0	79	67.1	75.0	62.5	70	50.0

TABLE 4. Comparison between UFLA with COBWEB/3, ECOBWEB and SBAC for heart dataset.

Algorithm	COBWEB/3	ECOBWEB	SBAC	UFLA
Accuracy	80.7	75.4	75.1	81.5
Confusion matrix	114 33 25 131	119 59 20 105	102 38 37 126	103 21 36 143



(a) (b)

FIGURE 2. VAT image of UFL distance matrix of the petroleum dataset: (a) before organizing and (b) after organizing. It shows that the data forms two clusters as obvious dark squares along the main diagonal.

2) DEFINE THE NUMBER OF CLUSTERS

VAT can facilitate the estimation of how many clusters exist by allowing the user to count the number of black squares along the diagonal. Fig. 2 shows the rearranged distance matrix resulting from the petroleum dataset according to the VAT algorithm. It clearly indicates two blocks of dark squares along the diagonal of the dissimilarity matrix. It is evident that the petroleum dataset has two clusters.

After the K-means step in Algorithm is performed with $k = 2$, the dataset is clustered two groups, a group of 400 and a group of 60.

3) CLUSTER ANALYSIS

To understand more about the partition structure, the distribution of features in each cluster are computed. The statistics of two clusters in the final results are computed and shown in Tables 5 and 6 for both categorical and numerical features.

Table 5 shows the distribution of the Formation Type attribute. Each of the two clusters contained projects from a different formation type; the projects in Cluster 1 were all from sandstone formations, while those in Cluster 2 were from unconsolidated formations. So the formation type feature has a significant discrimination information in the partition structure.

TABLE 5. Distribution of formation type feature in the two clusters.

	San	Dol	Unc	Tri	Lim	Con	Sha
C1	400	0	0	0	0	2	0
C2	0	3	50	1	2	3	1

TABLE 6. Distribution of numerical features in the two clusters.

	C1				C2			
	min	max	ave	std	min	max	ave	std
porosity	18	40	32	3.8	7.5	65	33	8.5
permeability	19	2e4	2260	1973	1	1.5e4	3839	4346
depth	100	9000	1491	829	175	5740	1594	1073
API	7	33	13	3.1	8	30	14.8	5.4
viscosity	10	8e5	1e4	5e+4	10	5e6	9e4	6e5
temperature	10	250	102	28	45	400	118	67
oil res.	29	100	67	15	32	100	231	15.2

Table 6 shows the statistics regarding the numerical attributes of the two clusters, revealing that many of the attributes yielded compact clusters. For example, for porosity, the deviation was only 3.8 and 8.5, while the values ranged from 18 to 40 and 7.6 to 65 for each of the two clusters.

On the other hand, attributes that had a large dynamic range still yielded a large deviation. For example, the deviation for viscosity in the two clusters was 5×10^4 and 6×10^5, respectively. From a dimensionality reduction point of view, features with large variations tend to contribute less significant information to the clustering process [41].

A closer study of the two cluster statistics reveals that permeability and temperature are strong indicators for clustering information since cluster 1 has a lower average and standard deviation but higher number of samples than cluster 2.

V. CONCLUSION

A novel methodology was presented based on UFL that works with noisy, uncertain and mixed data. For mixed-data applications, UFLA was presented. UFLA can learn its features even when the amount of data is small in important subspaces of the dataset. The learned feature representations can remove the distinction in treating categorical and numeric features, leading to a better clustering result. Visual assessment tendency is used to determine the true number of clusters in the dataset when the number of clusters is unknown. Results from the application of this method to several real datasets demonstrate the effectiveness of the approach.

REFERENCES

[1] R. Xu and D. Wunsch, *Clustering*. New York, NY, USA: Wiley, 2009.
[2] U. von Luxburg, "A tutorial on spectral clustering," *Statist. Comput.*, vol. 17, no. 4, pp. 395–416, 2007.
[3] V. Ganti, J. Gehrke, and R. Ramakrishnan, "CACTUS—Clustering categorical data using summaries," in *Proc. 5th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 1999, pp. 73–83.
[4] S. Guha, R. Rastogi, and K. Shim, "Rock: A robust clustering algorithm for categorical attributes," *Inf. Syst.*, vol. 25, no. 5, pp. 345–366, 2000.
[5] C. Li and G. Biswas, "Unsupervised learning with mixed numeric and nominal data," *IEEE Trans. Knowl. Data Eng.*, vol. 14, no. 4, pp. 673–690, Jul./Aug. 2002.
[6] Z. Huang, "Clustering large data sets with mixed numeric and categorical values," in *Proc. 1st Pacific-Asia Conf. Knowl. Discovery Data Mining (PAKDD)*, 1997, pp. 21–34.
[7] Z. Huang, "Extensions to the *k*-means algorithm for clustering large data sets with categorical values," *Data Mining Knowl. Discovery*, vol. 2, no. 3, pp. 283–304, 1998.
[8] J. Ji, W. Pang, C. Zhou, X. Han, and Z. Wang, "A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data," *Knowl.-Based Syst.*, vol. 30, pp. 129–135, Jun. 2012.
[9] Z. He, X. Xu, and S. Deng. (2005). "Clustering mixed numeric and categorical data: A cluster ensemble approach." [Online]. Available: http://arxiv.org/abs/cs/0509011
[10] C.-C. Hsu and S.-H. Wang, "An integrated framework for visualized and exploratory pattern discovery in mixed data," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 2, pp. 161–173, Feb. 2006.
[11] C.-C. Hsu and Y.-C. Chen, "Mining of mixed data with application to catalog marketing," *Expert Syst. Appl.*, vol. 32, no. 1, pp. 12–23, 2007.
[12] T.-W. Ryu and C. F. Eick, "Similarity measures for multi-valued attributes for database clustering," in *Proc. Smart Eng. Syst. Design Neural Netw., Fuzzy Logic, Evol. Program., Data Mining Rough Sets (ANNIE)*, 1998, pp. 25–29.
[13] M. Chavent, F. de A. T. de Carvalho, Y. Lechevallier, and R. Verde, "New clustering methods for interval data," *Comput. Statist.*, vol. 21, no. 2, pp. 211–229, 2006.
[14] G. A. Carpenter, S. Grossberg, and D. B. Rosen, "Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system," *Neural Netw.*, vol. 4, no. 6, pp. 759–771, 1991.
[15] J. C. Bezdek, R. J. Hathaway, and J. M. Huband, "Visual assessment of clustering tendency for rectangular dissimilarity matrices," *IEEE Trans. Fuzzy Syst.*, vol. 15, no. 5, pp. 890–903, Oct. 2007.
[16] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.
[17] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: Transfer learning from unlabeled data," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 759–766.
[18] J. C. van Gemert, J.-M. Geusebroek, C. J. Veenman, and A. W. M. Smeulders, "Kernel codebooks for scene categorization," in *Computer Vision*. Berlin, Germany: Springer-Verlag, 2008, pp. 696–709.
[19] L.-J. Li, H. Su, L. Fei-Fei, and E. P. Xing, "Object bank: A high-level image representation for scene classification & semantic feature sparsification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1378–1386.
[20] A.-H. Tan, G. A. Carpenter, and S. Grossberg, "Intelligence through interaction: Towards a unified theory for learning," in *Advances in Neural Networks*. Berlin, Germany: Springer-Verlag, 2007, pp. 1094–1103.
[21] L. Meng, A.-H. Tan, and D. Xu, "Semi-supervised heterogeneous fusion for multimedia data co-clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 9, pp. 2293–2306, Sep. 2014.
[22] A. Coates, A. Y. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, 2011, pp. 215–223.
[23] K. Bache and M. Lichman. (2013). *UCI Machine Learning Repository*. [Online]. Available: http://archive.ics.uci.edu/ml
[24] *Enhanced Oil Recovery (EOR) Survey Published Biannually by Oil and Gas Journal From 1980 ∼2012*. [Online]. Available: http://www.ogj.com/index.html
[25] M. Ranzato, Y.-L. Boureau, and Y. LeCun, "Sparse feature learning for deep belief networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 1185–1192.
[26] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 3361–3368.
[27] A. M. Cheriyadat, "Unsupervised feature learning for aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 439–451, Jan. 2014.
[28] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, vol. 2. Sep. 1999, pp. 1150–1157.
[29] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1. Jun. 2005, pp. 886–893.
[30] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," in *Proc. NIPS Workshop Deep Learn. Unsupervised Feature Learn.*, Granada, Spain, 2011, p. 5.
[31] M. Ranzato, F. J. Huang, Y.-L. Boureau, and Y. LeCun, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2007, pp. 1–8.
[32] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 19. 2007, pp. 153–160.
[33] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, "Why does unsupervised pre-training help deep learning?" *J. Mach. Learn. Res.*, vol. 11, pp. 625–660, Feb. 2010.
[34] H. Larochelle, Y. Bengio, J. Louradour, and P. Lamblin, "Exploring strategies for training deep neural networks," *J. Mach. Learn. Res.*, vol. 10, pp. 1–40, Jan. 2009.
[35] D. S. Levine, *Introduction to Neural and Cognitive Modeling*. New York, NY, USA: Psychology Press, 2000.
[36] A.-H. Tan, H.-L. Ong, H. Pan, J. Ng, and Q.-X. Li, "Towards personalised Web intelligence," *Knowl. Inf. Syst.*, vol. 6, no. 5, pp. 595–616, 2004.
[37] T. C. Havens and J. C. Bezdek, "An efficient formulation of the improved visual assessment of cluster tendency (iVAT) algorithm," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 5, pp. 813–822, May 2012.
[38] H.-S. Park and C.-H. Jun, "A simple and fast algorithm for K-medoids clustering," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 3336–3341, 2009.
[39] K. McKusick and K. Thompson, "COBWEB/3: A portable implementation," NASA, Moffet Field, CA, USA, Tech. Rep. FIA-90-6-13-2, 1990.
[40] Y. Reich and S. J. Fenves, "The formation and use of abstract concepts in design," in *Concept Formation Knowledge and Experience in Unsupervised Learning*. San Mateo, CA, USA: Morgan Kaufmann, 1991.
[41] A. Zimek and J. Vreeken, "The blind men and the elephant: On meeting the problem of multiple truths in data from clustering and pattern mining perspectives," *Mach. Learn.*, vol. 98, no. 1, pp. 121–155, 2015.

DAO LAM received the B.S. degree in computer engineering from the Post and Telecommunications Institute of Technology, Ho Chi Minh City, Vietnam, in 2003, the M.S. degree in computer engineering from Waseda University, Japan, in 2008, and the Ph.D. degree in computer engineering from the Missouri University of Science and Technology, in 2015. His research interests are machine learning, computer vision, and robotics. He was a recipient of the Japanese Government Scholarship in 2006.

MINGZHEN WEI received the bachelor's and master's degrees in petroleum engineering from the China University of Petroleum, and the Ph.D. degree in computer science from the New Mexico Institute of Mining and Technology. She is currently an Assistant Professor with the Petroleum Engineering Program, Department of Geosciences and Geological and Petroleum Engineering, Missouri University of Science and Technology. Her research mainly focuses on reservoir simulation, unconventional resources, and application of artificial intelligence on the oil and gas problems.

DONALD WUNSCH (F'05) received the B.S. degree in applied mathematics from the University of New Mexico, and completed the Jesuit Core Honors Program, Seattle University, the M.S. degree in applied mathematics and the Ph.D. degree in electrical engineering from the University of Washington, Seattle, and the E.M.B.A. degree from Washington University in St. Louis. He was with Texas Tech University, Boeing, Rockwell International, and International Laser Systems. He is currently the Mary K. Finley Missouri Distinguished Professor with Missouri University of Science and Technology (Missouri S&T). His key research contributions are clustering; adaptive resonance and reinforcement learning architectures, hardware, and applications; neurofuzzy regression; traveling salesman problem heuristics; robotic swarms; and bioinformatics. His research has been cited over 10 000 times. He has produced 17 Ph.D. recipients in computer engineering, electrical engineering, and computer science. He has attracted over U.S. $8 million in sponsored research. He has authored over 400 publications, including nine books. He was the INNS President, an INNS Fellow, and the Senior Fellow from 2007 to 2013. He was a winner of the NSF CAREER Award and the 2015 INNS Gabor Award. He served as the IJCNN General Chair, and on several Boards, including the St. Patrick's School Board, the IEEE Neural Networks Council, the International Neural Networks Society, and the University of Missouri Bioinformatics Consortium. He chaired the Missouri S&T Information Technology and Computing Committee and the Student Design and Experiential Learning Center Board.