IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

A Survey of Mixed Data Clustering Algorithms

AMIR AHMAD[1] and SHEHROZ KHAN [2],
[1]College of Information Technology, United Arab Emirates University,Al-Ain, UAE (e-mail: amirahmad@uaeu.ac.ae)
[2]Toronto Rehabilatatin Institute, Universiy Health Network, 550, University Avenue, Toronto, Canada, (e-mail: shehroz.khan@utoronto.ca)
Corresponding author: AMIR AHMAD (e-mail: amirahmad@uaeu.ac.ae).

"This work was supported by a UAE university Start-up grant (grant number G00002668; fund number 31T101."

ABSTRACT Most of the datasets normally contain either numeric or categorical features. Mixed data comprises of both numeric and categorical features, and they frequently occur in various domains, such as health, finance, marketing, etc. Clustering is often sought on mixed data to find structures and to group similar objects. However, clustering mixed data is challenging because it is difficult to directly apply mathematical operations, such as summation, average etc. on the feature values of these datasets. In this paper, we review various types of mixed data clustering techniques in detail. We present a taxonomy to identify ten types of different mixed data clustering techniques. We also compare the performance of several mixed data clustering methods on publicly available datasets. The paper further identifies challenges in developing different mixed data clustering algorithms and provides guidelines for future directions in this area.

INDEX TERMS Categorical Features, Clustering, Mixed Datasets, Numeric Features

I. INTRODUCTION

Most of the datasets contain numeric and categorical features [13], [175]. Numeric features can take real values, such as height, weight, distance, etc. Categorical features can take fixed number of values, such as color, type of job, blood group of a person, etc. Clustering is often performed to group datasets, such that data points in a group are similar to each other and dissimilar from those in other groups based on some notion of a similarity measure [87]. To compute the similarity between feature values, mathematical operations (such as summation, mean, etc.) are applied on them. These mathematical operations can be easily applied on numeric features; however, they cannot be applied directly on categorical features. Hence, computing similarity for categorical data points is a challenging task [16]. Some methods have been suggested for computing the similarity of data points having categorical features [16]. Many datasets contains both numeric and categorical features; they are called *mixed datasets*. The problem of computing similarity between two data points becomes more difficult if the dataset has both types of features. An example of mixed dataset is given in Table 1. The dataset has four features; Height and Weight are numeric features whereas Blood Group and Profession are categorical features. A simple strategy to find similarity between two data points in this dataset is to split the numerical and categorical part. Then, find the Euclidean distance between two data points for numerical features and Hamming distance for the categorical features [78]. However, it is not clear if both the distance measures calculate 'similar' type of similarity and if the scales of these distances are same or not. It is also not clear the proportions in which both the distance measures are combined. Therefore, until the notion of similarity is not clarified for the mixed data, clustering on it will be even more challenging.

TABLE 1. A mixed dataset.

Weight (Kg.)	Height (Meter)	Blood Group	Profession
80.6	1.85	B+	Teaching
73.6	1.72	A+	Teaching
70.8	1.79	B+	Medical
85.9	1.91	A-	Sportsman
83.4	1.65	A+	Medical

Clustering of mixed data has been applied in various domains such as marketing, medical, finance etc. [3], [34], [84]. Many review papers have been published that covers different types of clustering algorithms, such as general clustering algorithms [1], [86], [176], evolutionary clustering algorithms [69], kernel and spectral clustering algorithms [50], cluster ensembles [160], and subspace clustering [141]. Some research articles also discuss mixed data clustering algorithms albeit partially [1]. Many clustering algorithms have also been proposed for mixed datasets in the past [77], [78], [133]. However, there has been no review paper that discussed mixed data clustering algorithms in detail. Recently there is a short review paper on mixed data clustering [8]. Many important mixed data clustering algorithms are not discussed in the paper [4], [15], [31], [32], [35], [37], [54], [60], [66], [67], [92], [108], [112], [133], [139], [146], [147], [161], [165]. The paper also does not discuss the challenges and future directions in this area in detail. It is important from the machine learning perspective to understand the challenges in this domain, evaluate the competing algorithms and apply in novel application areas.

In this paper, we present a comprehensive review of clustering algorithms for mixed datasets. We present a taxonomy to identify ten different types of mixed data clustering algorithms based on the methodology used to cluster mixed data points. Then, we delve into each type of mixed data clustering and analyze different algorithms, their functioning, strengths and weaknesses. We also present experimental results of many mixed data clustering algorithms and compare their performance over several publicly available datasets. This step is important to note the relative importance and advantages of different mixed data clustering algorithms. We identify challenges among different types of mixed data clustering algorithms and highlight opportunities in this research area.

The paper is organized in the following manner. In Section II-A, we present a taxonomy to identify different types of mixed-data clustering algorithms. Section III discusses about the most cited mixed data clustering algorithms along with their performances. Section IV highlights several publicly available tools for performing mixed data clustering. Applications of mixed data clustering to various domains are also discussed in this section. In Section V, we discuss the challenges and opportunities for future work in this area. We conclude our paper in Section VI.

II. MIXED DATA CLUSTERING ALGORITHMS

Mixed data clustering can be done in several ways based on the process involved in clustering of data points. Based on our review of literature, in this section, we present a taxonomy to identify different types of mixed data clustering algorithms. Table 2 shows ten different groups of mixed data clustering methods, along with their names and relevant papers we will discuss next. In this section, we will review the literature around these groups.

A. PARTITIONAL ALGORITHMS

Partitional clustering algorithms consider the center of data points as the center of the corresponding cluster [87]. K-means clustering algorithm [119] is a partitional clustering algorithm. It is linear in the number of data points and scales well for large datasets. The algorithm minimizes the following cost function iteratively,

$$\sum_{i=1}^{n} \xi(d_i, C_i) \tag{1}$$

where n is the number of data points in the dataset, C_i is the nearest cluster center of data point d_i. ξ is a chosen distance measure between d_i and C_i.

In K-means clustering, the center of a cluster is computed by calculating the mean of the data points in that cluster. Generally, Euclidean distances between a data point and different cluster centers are computed to determine the nearest center for each data point, and the data point is assigned to that cluster. However, Euclidean distances and mean values cannot be calculated for mixed datasets. To deal with this situation, several variants of K-means clustering algorithms have been proposed to deal with mixed datasets.

Huang [77], [78] present K-prototypes clustering algorithm for mixed datasets that proposes a new cost function. New representations of cluster centers and a new definition of distance between a data point and a cluster center are proposed for mixed datasets. Cluster centers are represented by mean values for numeric features and mode values for categorical features. However, the proposed cluster center does not represent the underlying clusters well, because (i) the mode for categorical features incurs loss of information, and (ii) the Hamming distance [16] is not a good representative of similarity between feature values for a pair of multi-valued categorical feature values. Ahmad and Dey [3] propose a new cost function and a distance measure to address these problems. In this method, they calculate the similarity between two feature values of a categorical feature from the data. The similarity depends upon co-occurrence of these feature values with feature values of other features. The weights of numeric features are also calculated in this method such that more significant features get more weights. A novel frequency based representation of cluster center is proposed for categorical features [3]. It is shown that their proposed clustering algorithm performs better than K-prototypes clustering algorithm [3].

TABLE 2. Types of mixed data clustering algorithms.

#	Types of clustering Algorithms	References
1	Partitional algorithms	[3], [76]–[78], [89], [95], [133], [153], [157], [172], [174], [182]
2	Hierarchical clustering algorithms	[49], [71], [74], [75], [94], [106], [145], [164]
3	Model Based clustering algorithms	[7], [26], [36], [48], [52], [81], [105], [123], [126], [151], [159], [169]
4	Neural networks-based clustering	[39], [41], [70], [72], [73], [75], [104], [136], [140], [167]
5	Other types of clustering algorithms	[2], [4], [15], [31], [32], [35], [37], [42]–[44], [54], [60], [66], [67], [91], [92], [102], [108], [112], [114], [139], [143], [146], [147], [161], [165]

Huang et al. [76] extend K-prototypes clustering algorithm to propose W-K-prototypes clustering algorithm in which, in each iteration the feature weights are computed and used in the cost function. These weights are inversely proportional to the sum of the within cluster distances. Their results suggest an improvement in clustering results with feature weights over the clustering results achieved with K-prototypes algorithm [77], [78]. Zao et al. [182] use the frequency of feature values for categorical features to define the cluster centers. Hamming distance measure was used to compute the distance for categorical features. Mean values are used for numeric features. They show better clustering results in comparison to K-prototypes algorithm [77], [78].

Modha and Spangler [133] employ weighting in K-means clustering. In this method, each data point is represented in different types of feature spaces. A measure is proposed to compute the distortion between two data points in each feature space. The distortions in different feature spaces are combined to compute feature weights. The method is also employed for mixed data clustering. A mixed dataset is considered having two feature spaces; one consisting of numerical features and the other categorical features. Each numerical feature is linearly scaled (a feature value is subtracted by the mean and divided by standard deviation) and 1-in-q representation for each q-ary categorical feature is used. Squared Euclidean distance is used for numeric features whereas cosine distance is used for categorical features.

Chen and He [27] use the distance measure suggested by Ahmad and Dey [3] to propose a mixed data clustering algorithm for data streams with mixed numerical and categorical features. The concept of micro-clusters is used in the algorithm. Micro-clusters are used to compress the data efficiently in data streams. In the first stage, initial cluster centers are calculated that are used to cluster the data. The method uses two parameters: decay factor and dense threshold. Decay factor defines the significance of historical data to current cluster whereas dense threshold is used in distinguishing dense mirco-clusters and sparse micro-clusters. The parameter optimization is a problem with the method.

Ran et al. [153] use the cluster centers proposed by Ahmad and Dey [3] for a novel mixed data clustering algorithm. Euclidean distance for numeric features and Hamming distance for categorical features with Gaussian kernel function is used to compute the distance between the cluster center and a data point. Ji et al. [89] combine the definition of cluster center [3] with the significance of feature term [76] to propose a new cost function. The significance of a feature is initially selected randomly, followed by update in values with each iteration. The random selection of the significance of a feature can make the random initialization of cluster center problem [98], [99] worse.

Roy and Sharma [157] extend fast genetic K-means cluster technique (FGKA) [115] for mixed data. The algorithm minimizes the total within-cluster variation. They use the distance measure proposed by Ahmad and Dey [3] in their algorithm. Chiodi et al. [33] propose an iterative partitional clustering algorithm for mixed data, which is motivated by the K-means clustering algorithm [120]. The Euclidean distance measure is used for a numeric feature and Hamming distance measure is used for categorical features. Mean values are used for numeric features and the frequency distribution of categorical values in clusters. Kacem et al. [95] proposes parallelization of K-prototypes clustering method [78] to handle big mixed datasets. The algorithm uses MapReduce framework [38] for parallelization. In Table 3, we summarize different K-means clustering type algorithms for mixed data clustering.

The other approach which has also been used with K-means type clustering algorithm to cluster mixed datasets is to first convert mixed datasets into numeric datasets and then apply K-means clustering algorithm on the numeric datasets. Barcelo-Rico and Jose-Luis [9] develop a method that uses polar or spherical coordinates to codify categorical features into numeric features; then K-means clustering algorithm is used on the new numeric datasets. Wang et al. [172] propose the context-based coupled representation for mixed datasets. The interdependence of numeric features and the interdependence of categorical features are computed separately. Then the interdependence across the numeric features and categorical features are computed. These relationships form the numeric representation for mixed-type data points. K-means clustering algorithm is used to cluster these new data points. Their experimental results suggest that the method outperform other mixed-data clustering algorithms. Wei et al. [174] propose a mutual information-based transformation method for unsupervised features that can convert categorical features into numeric features, which is then clustered by using K-means clustering algorithm. Lam et al. [104] use unsupervised feature learning approach to get sparse representation of mixed datasets. Fuzzy adaptive resonance theory (ART) approach [23] is used to create new features. Firstly, fuzzy ART approach is used to create prototypes of the dataset, which are employed as mixed features encoder to map individual data points in the new feature space. They use K-means clustering algorithm to cluster data points in the new feature space. Table 4 summarizes those methods that first convert the mixed data into numeric data then apply the K-means type clustering techniques on the new numeric data.

The traditional K-means and K-modes algorithms suffer from several drawbacks, such as cluster center initialization [98], [99] and the prior knowledge of the number of cluster [119]. These issues also exist in the K-means clustering type algorithms for clustering mixed datasets, due to their conceptual similarity. In the next sub-section, we review relevant literature that covers these mentioned issues.

1) Cluster center initialization

Cluster center initialization is a well known problem with the K-means type clustering algorithms [98], [99]. In these algorithms, generally initial cluster centers are selected randomly. Consistent results may not be achieved in different runs of the algorithm due to random selection of initial cluster

TABLE 3. K-means clustering type algorithm for mixed datasets.

Algorithm	Center Definition	Distance Measure
Huang [77], [78]	Mean values for numeric features, mode values for categorical data	Euclidean distance for numeric features, Hamming distance for categorical features
Ahmad and Dey [3]	Mean values for numeric features, proportional frequency based center for categorical features	Weights for numeric features are calculated, Euclidean distance for numeric features and co-occurrence based distance measure for categorical features
Huang et al. [76]	Mean values for numeric features, mode values for categorical features	Weights of features based on the importance of the features in clustering are calculated in each run with distance measure used by Huang [77], [78]
Zhao et al. [182]	Mean values for numeric features, proportional frequency based center for categorical faetures	Euclidean distance for numeric features, Hamming distance for categorical features
Modha and Spangler [133]	First, 1-in-q representation for each q-ary categorical feature, Mean values for all features	Weights of features are calculated, squared Euclidean distance is used for numeric features whereas cosine distance is used for categorical features
Ji et al. [89]	center as proposed by Ahmad and Dey [3]	Weights are calculated by the method suggested by Huang et al. [76], squared Euclidean distance is used for numeric features, Hamming distance is used for categorical features
Ran et al. [153]	center as proposed by Ahmad and Dey [3]	Gauss kernel function

TABLE 4. Clustering algorithm when categorical features are converted to numeric features.

Algorithm	Method to convert the categorical features to numeric features
Barcelo-Rico and Jose-Luis [9]	Coding is based on polar or spherical coordinates
Wang et al. [172]	Context based coupled relationship for mixed data
Wei et al. [174]	Mutual information (MI)-based unsupervised feature transformation
Lam et al. [104]	Fuzzy adaptive resonance theory [23]

centers; thus data mining researchers may find it problematic to analyze such clustering results. Ji et al. [90] suggest an algorithm to create initial cluster centers for K-means type clustering algorithms for mixed datasets. They introduce an idea of the centrality of data points that uses the concept of neighbor-set. The centrality and distances are used to compute initial cluster centers. However, their algorithm has quadratic complexity that contravenes the linear time complexity benefit of the K-means clustering type algorithms.

Using density peaks [156], Chen et al. [28] propose a novel algorithm to determine the initial cluster centers for mixed datasets. Higher density points are used to identify cluster centers. This algorithm has quadratic complexity, hence, it is not useful for K-means clustering type algorithms. Wangchamhan et al. [173] combine a search algorithm, League Championship Algorithm [96], with K-means clustering algorithm to identify the initial cluster centers. They apply Gower's distance measure [59] to find the distance between a data point and a cluster center.

Ahmad and Hashmi [64] combine the distance measure and the definition of centers for mixed data proposed by Ahmad and Dey [3] with the cost function of K-harmonic clustering [20] to extend the K-harmonic clustering for numeric data to mixed data. Their results suggest that the suggested method is quite robust to the selection of initial cluster centers as compared to other K-means clustering type algorithms for mixed datasets. Zheng et a. [183] combine evolutionary algorithm (EA) with K-prototypes clustering algorithm [78]. The global searching ability of EA makes the proposed algorithm less sensitive to cluster initialization.

2) Number of clusters

In K-means type clustering algorithms, the number of cluster is user defined, which may not be a true representative of the natural number of clusters in the data. Liang et al. [110] propose a cluster validity index to find out the number of clusters for mixed data clustering. This cluster validity index has two components; one for numeric features and the other for categorical features. For categorical features, the cluster validity index uses the category utility function developed by Gluck and Corter [117]. Whereas, for numeric features, a corresponding category utility function proposed by Mirkin [17] is used. Each component is given a weight depending upon the number of categorical and numeric features and the total number of features. This cluster validity index is computed for different number of clusters. The number of clusters which produce the maximum value of cluster validity index is chosen as the optimal number of cluster. In this method, the process starts with a large number of clusters and in each round the worst cluster is combined with other clusters. Renyi entropy [154] for numeric features and complement entropy [109] for categorical features are used to determine the worst cluster. The method is used with K-prototypes method [78].

Rahman and Islam [150] combine genetic algorithm optimization technique [132] and K-means clustering algorithm to produce a clustering algorithm for mixed data that computes the number of clusters automatically. They use the distance measure proposed by Rahman and Islam [118] to compute the distance between a pair of categorical values. The algorithm shows good results; however, its complexity is quadratic.

B. HIERARCHICAL CLUSTERING

Hierarchical clustering is a clustering method that creates a hierarchy of clusters by using a similarity matrix [87]. The similarity matrix is constructed by determining the similarity of each pair of data points. These algorithms have a large time complexity (generally $O(n^3)$ [87]).

Philip and Ottaway [145] use Gower's similarity measure [59] to compute the similarity matrix for mixed datasets. Gower's similarity measure computes the similarity by dividing features into two subsets one for categorical features and the other for numeric features. Hamming distance is applied to compute the similarity between two points of data points for a categorical feature. A weighted average of similarities for all categorical features is the similarity between two data points in a categorical feature space. For numeric features, the similarity is computed such that the same values give the similarity value 1, whereas if the difference between the values is maximum possible difference (the difference between maximum and minimum values of the feature) the similarity is 0. The sum of similarity values for all the numeric features is the similarity for two data points in a numeric feature space. The similarity of the categorical feature space and the numeric feature space are added to compute the similarity between two data points. Then hierarchical agglomerative clustering is used to create clusters.

Chiu et al. [49] develop a similarity measure to compute the similarity between two clusters for mixed data. This dissimilarity measure is related with the decrease in log-likelihood function when two clusters are merged. The authors combine the Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) clustering algorithm [181] which uses hierarchical clustering algorithm with their proposed similarity measure to develop a clustering algorithm that can handle mixed datasets.

Li and Biswas [106] propose Similarity-Based Agglomerative Clustering (SBAC) algorithm for mixed data clustering. SBAC uses Goodall similarity measure [58] and applies a hierarchical agglomerative approach to build cluster hierarchies. Goodall similarity approach gives more weight to uncommon feature value matches. It applies group-average methods [87] to perform the aggregation process.

Hsu et al. [74] propose a distance measure based on a distance hierarchy that uses a concept hierarchy [62], [63], which is consists of concept nodes and links. The more general concepts are represented by higher-level nodes, whereas more specific concepts are represented by lower-level nodes. The categorical values are represented by a tree structure such that each leaf of a tree is represented by a categorical value. Each feature of a data point is associated with a distance hierarchy. The distances between two data points is calculated by using their associated distance hierarchies. An agglomerative hierarchical clustering algorithms [87] is applied to a dissimilarity matrix to compute the clusters. Domain knowledge is required to make distance hierarchies for categorical features which is not simple in many cases. Hsu and Chen [71] propose a new similarity measure to cluster mixed data. The algorithm uses variance for computing the similarity of numeric values. For computing the similarity between categorical values, they [71] utilizes entropy with distance hierarchies [74]. The similarities are then aggregated to compute the similarity matrix for a mixed dataset. Incremental clustering is used on the similarity matrix to compute the clusters. Hsu and Huang [75] extend this work [71] by applying ART to cluster data points by using the distance hierarchies as the input of the network. Shih et al. [164] convert categorical features of a mixed dataset into numeric features by using frequencies of co-occurrence of categorical feature values. Then, the dataset with all numeric features is clustered by using hierarchical agglomerative clustering algorithm [87].

Lim et al. [111] partition the data into two parts; the categorical data and numeric data. Both data are clustered separately. The clustering results are combined by using a weighted scheme to get a similarity matrix. Agglomerative hierarchical clustering method is applied on the similarity matrix to get the final clusters. Gower's similarity measure assign equal weight to both types of feature in computing the similarity between two data points. The similarity matrices may be dominated by one kind of feature type. Chae et al. [94] assign weights to the different feature types to overcome this problem. Improved clustering results are shown with these weighted similarity matrices. Different mixed data clustering algorithms based on hierarchical clustering that are discussed in this section are summarized in Table 5.

C. MODEL BASED CLUSTERING

Model based clustering uses an assumption that a data point matches a model, that, in many cases, is a statistical distribution [129]. The model is generally user defined, it may give poor results if a proper model is not selected. Model based clustering algorithms are generally slow than K-means type clustering algorithms [129].

AUTOCLASS [26] performs clustering by integration of finite mixture distribution and Bayesian methods with prior distribution of each feature. Autoclass can cluster data containing both categorical and numeric features. Everitt [48] proposes a clustering algorithm by using model-based clustering for datasets consisting of both numeric features and binary or ordinal features. The normal model is extended to handle mixed datasets by using thresholds for the categorical features. Due to high computational cost, the method is only useful for datasets containing very few categorical features. To overcome this problem, Lawrence and Krzanowski [105] extend homogeneous Conditional Gaussian model to the finite mixture case, to compute maximum likelihood estimates for for the parameters in a sample population. They suggest that their method works for arbitrary number of features.

Moustaki and Papageorgiou [135] use a latent class mix-

TABLE 5. Hierarchical clustering algorithms for mixed datasets.

Algorithm	Similarity measure for a similarity matrix	Clustering algorithm
Philip and Ottaway [145]	Gower's similarity Matrix [59]	Agglomerative hierarchical clustering method
Chiu et al. [49]	Probabilistic model by using a log-likelihood function	BIRCH algorithm [181]
Li and Biswas [106]	Goodall similarity measure [58]	Agglomerative hierarchical clustering with group- average method
Hsu et al. [74]	Distance hierarchy by using concept hierarchy [62], [63]	Agglomerative hierarchical clustering
Hsu and Chen [71]	Variance for numeric features and entropy with distance hierarchies [74] for categorical features	Incremental clustering
Hsu and Huang [75]	Similarity measure proposed by Hsu and Chen [71]	Adaptive resonance theory network [22]
Shih et al. [164]	Convert categorical features into numeric features	Hierarchical agglomerative clustering algorithm [87]
Lim et al. [111]	Two similarity matrix, one for categorical data and one for numeric data	Agglomerative hierarchical clustering method
Chae et al. [94]	Modified Gower's similarity matrix	Agglomerative hierarchical clustering method

ture model for mixed data clustering. Browne and McNicholas [18] propose a mixture of latent features model for clustering, the expectation-maximization (EM) framework [40] is used for model fitting. Andreopoulos et al. present [7] a clustering algorithm, Bi-Level Clustering of Mixed categorical and numerical data types (BILCOM) for mixed datasets. The algorithm uses categorical data clustering to guide the clustering of numerical data. Hunt and Jorgensen [81]–[83] propose a mixture model clustering approach for mixed data. In this approach, a finite mixture of multivariate distributions is fitted to data and then the membership of each data point is calculated by computing the conditional probabilities of cluster membership. Local independence assumption can be used to reduce the model parameters. They further show that the method can also be applied for clustering mixed datasets with missing values [82].

ClustMD method [126] uses a latent variable model to cluster mixed datasets. It is suggested that a latent variable, with a mixture of Gaussian distributions, produces the observed mixed data. An EM algorithm is applied to estimate parameters for ClustMD. Monte Carlo EM algorithm [125] is used for datasets having categorical features. This method can model both the numeric and categorical features; however, it becomes computationally expensive as the number of features increase. To overcome this problem, McParland et al. [127] propose a clustering algorithm for high dimensional mixed data by using a Bayesian finite mixture model. In this algorithm, the estimation is done by using Gibbs sampling algorithm. To select the optimal model, they also propose approximate Bayesian Information Criterion-Markov chain Monte Carlo criterion. They show that the method works well on a mixed medical data consisting of high dimensional numeric phenotypic features and categorical genotypic features. Saadaoui et al. [159] propose a projection of the categorical features on the subspaces spanned by numeric features. Then an optimal Gaussian Mixture Model is obtained from the resulting Principal Component Analysis (PCA)-regressed subspaces.

Copulas are defined as "functions that join or couple multivariate distribution functions to their one-dimensional marginal distribution functions" and as "distribution functions whose one-dimensional margins are uniform." [137]. Rajan and Bhattacharya [151] present a clustering algorithm based on Gaussian mixture copula that can model dependencies between features and can be applied for datasets having numeric and categorical features. Their method outperforms other clustering algorithms on a variety of datasets. Tekumalla et al. [169] use the concept of vines for mixed data clustering, wherein they propose an inferencing algorithm to fit those vines on the mixed data. A dependency-seeking multi-view clustering that uses Dirichlet process mixture of vines is developed [169]. Marbac et al. [123] present a mixture model of Gaussian copulas for mixed data clustering. In this model, a component of the Gaussian copula mixture creates a correlation coefficient for a pair of features. They selected the model by using two information criteria: Bayesian information criterion [162] and integrated completed likelihood criterion [11]. The Bayesian inference is performed by using a Metropolis-within-Gibbs sampler. Foss et al [52] develop a semi-parametric method, KAMILA (KAy-means for MIxed LArge data), for clustering mixed data. KAMILA balances the effect of the numeric and categorical features on clustering. KAMILA integrates two different kinds of clustering algorithms; the K-means algorithm and Gaussian-multinomial mixture models [83]. Similar to K-means clustering algorithm, no strong parametric assumptions are made for numeric features in KAMILA algorithm. KAMILA uses the properties of Gaussian-multinomial mixture models to balance the effect of numeric and categorical features without specifying weights.

Table 6 summarizes various model based clustering algorithms for mixed data that are discussed in this section.

D. NEURAL NETWORKS-BASED CLUSTERING

Self-organizing map (SOM) [100], [101] is a neural network that is used to nonlinearly project a dataset onto a

TABLE 6. Model based clustering algorithms for mixed datasets.

Algorithm	Model
AUTOCLASS [26]	Bayesian methods
Everitt [48]	Model-based clustering with the use of thresholds for the categorical features.
Lawrence and Krzanowski [105]	Extension of homogeneous conditional Gaussian model to the finite mixture situation.
Moustaki and Papageorgiou [135]	Latent class mixture model.
Browne and McNicholas [18]	A mixture of latent variables model with the expectation-maximization framework. [40].
BILCOM [7]	Pseudo-Bayesian process with categorical data clustering to guide the clustering of numerical data.
Hunt and Jorgensen [81]–[83]	A finite mixture of multivariate distributions is fitted to data.
ClustMD method [126]	A latent variable model.
McParland et al. [36]	Bayesian finite mixture model.
Saadaoui et al. [159]	A projection of the categorical features on the subspaces spanned by numeric features and then the application of Gaussian Mixture Model.
Rajan and Bhattacharya [151]	Gaussian mixture copula.
Tekumalla1 et al. [169]	Vine copulas and Dirichlet process mixture of vines.
Marbac [123]	A mixture model of Gaussian copulas.
KAMILA [52]	K-means algorithm and Gaussian-multinomial mixture models

lower-dimensional feature space so that clusters analysis can be done in the new feature space. Traditional SOM based clustering methods can handle numeric features, however it cannot be used directly for categorical features. Categorical attributes are first transformed into binary features which are treated as numeric features [41]. Hsu [70] develops Generalized SMO model to compute similarity of categorical values by using a distance hierarchy that is based on a concept hierarchy. It consists of nodes and weighted links; more general concepts are represented by higher level nodes whereas more specific concepts are represented by lower level nodes. Distance hierarchies are also used to compute the similarities between two data points in the complete feature space (numeric and categorical features). Visualization-Induced SMO [178] has better preservation of the structure of data in the new low dimensional space as compared to SMO. Hsu and Lin [72] combine Generalized SMO with Visualization-Induced SMO to develop a method Generalized visualization-Induced SOM to cluster mixed datasets. The experiments suggest that the method gives excellent cluster analysis results. Hsu and Lin [73] modify the distance measure presented in Generalized SMO and use the Visualization-Induced SMO to develop a new method for mixed data clustering. Traditional SMO has a weaknesses that it has predefined fixed-size map, to improve the flexibility of SMO, Growing SMO is proposed [5]. Growing SMO starts with a small size of map and grows with training data. Tai and Hsu [167] integrate Generalized SMO with Growing SMO to develop a clustering algorithm for mixed datasets. Chen and Marques [136] propose a clustering algorithm by using SMO, this method uses Hamming distance for categorical features and Euclidean distance for numeric features. This method gives more weight to categorical attributes, to overcome this problem Coso et al. [39] modify the distance measure such that each type of feature has equal weight. The method show better results than the method presented by Chen and Marques.

Noorbehbahani et al. [140] propose an incremental mixed-data clustering algorithm which uses Self-Organizing Incremental Neural Network algorithm [55]. They also propose a new distance measure in which the distance between two categorical values are dependent on the frequencies of these features. The co-occurrence of feature values [3] are not considered, which may affect the accuracy of the distance measure.

Lam et al. [104] uses Fuzzy adaptive resonance theory (ART) approach [23] to create new numeric features from the mixed features and then K-means clustering algorithm is used cluster data points in new feature space. Hsu and Huang [75] uses ART to create similarity matrix that can be used to cluster data points by using hierarchical clustering.

E. OTHER CLUSTERING ALGORITHMS FOR MIXED DATA

In this section, we will discuss several other types of mixed data clustering algorithms that may not fit in the scope of above discussed major categories.

1) Spectral clustering

Spectral clustering techniques [138] perform dimensionality reduction by using eigenvalues of the similarity matrix of the data. Thereafter, the clustering in performed in fewer dimensions. First a similarity matrix is computed then a spectral clustering algorithm [138] is applied on this similarity matrix to obtain clusters. Luo et al. [116] propose a similarity measure by using a clustering ensembles technique. In this measure, the similarity of two data points is computed separately for numeric features and categorical features. The two similarities are added to get the similarity between two data points. Niu et al. [139] present a clustering algorithm for mixed data, in which the similarity matrices for numeric features and categorical features are computed separately. Coupling relationships of features are used to compute similarity matrices. Then both matrices are combined by weighted summation to compute the similarity matrix for the mixed data. David and Averbuchb [37] propose a clustering algorithm, SpectralCAT, that uses categorical spectral clustering to cluster mixed datasets. The algorithm automatically transforms the numeric features into categorical values. It is performed by finding the optimal transformation according to the Calinski and Harabasz index [21]. Then, a spectral clustering method on the transformed data is applied [37].

2) Subspace clustering

Subspace clustering [142] seeks to discover clusters in different subspaces within a dataset. Ahmad and Dey [4] use a distance measure [3] for the mixed data with a cost function for subspace clustering [93] to develop a K-means type clustering type algorithm, which can produce subspace clustering of mixed data. Jia and Cheung [92] present a feature-weighted clustering model that uses data point-cluster similarity for soft subspace clustering of mixed datasets. They propose a unified weighting scheme for the numeric and categorical features, which determines the feature-to-cluster contribution. The method finds most appropriate number of clusters automatically. Plant and Bohm [147] develop a clustering technique, INCONCO, which produces interpretable clustering results for mixed data. The algorithm uses the concept of data compression by using the Minimum Description Length (MDL) principle [155]. INCONCO identifies the relevant feature dependencies using linear models and provides subspace clustering for mixed datasets. INCONCO does not support all types of feature dependencies.

3) Density based clustering algorithms

Density based clustering methods assume that clusters are defined by dense regions in the data space, separated by lower dense regions [47]. Du et al. [42], [44] propose a new distance measure for mixed data clustering. In this measure, they assign a weight to each categorical feature. They combine this distance measure with density peaks clustering algorithm [156] to cluster mixed datasets. However, the selection of different parameters makes it difficult to be used in practice. Liu et a. [114] propose a density based clustering algorithm for mixed datasets. The authors extend “density-based spatial clustering of applications with noise” (DBSCAN) [47] algorithm for mixed datasets. Entropy is used to compute the distance measure for mixed datasets. Milenova and Campos [131] use orthogonal projections to cluster mixed datasets. These orthogonal projections are used to find high density regions in the input data space.

4) Conceptual clustering

Conceptual clustering [51] generates concept description for each generated cluster. Generally, conceptual clustering methods generate hierarchical category structures. COBWEB [51] use Category Utility (CU) measure [117] to define the relation between groups or clusters. As CU measure can only handle categorical features, CU measure is extended to handle numeric features for mixed data clustering. COBWEB3 [124] integrates the original COBWEB algorithm with the methodology presented in CLASSIT [56] to deal with numeric features in the CU measure. In this method, it is assumed that numeric feature values are normally distributed. To overcome the problem of normal distribution assumption, a new method ECOBWEB [152] is presented. In this method, the probability distribution about the average for a feature is used.

5) Fuzzy clustering

Fuzzy clustering represent those approaches where a data point can belong to more than one cluster with different degree (or probability) of membership [177]. Various fuzzy clustering algorithms have been proposed for mixed data clustering [2], [43], [91]. Ahmad and Dey [2] use a dynamic probabilistic distance measure to determine the weights of numeric features and distances between categorical values for each pair of categorical values of a categorical feature. The distance measure is combined with the cluster center definition suggested by El-Sonbaty and Ismail [46] to develop a Fuzzy C-means (FCM) clustering type algorithm [10], [45] for mixed data. Ji et al. [91] propose a fuzzy clustering method for mixed datas by combining the similarity measure proposed by Ahmad and Dey [3] and the cluster center definition suggested by El-Sonbaty and Ismail [46].

Doring et al. [43] propose a fuzzy clustering algorithm for mixed data by using a mixture model. The mixture model is used to determine the similarity measure for mixed datasets. It also helps in the finding of cluster prototypes. The inverse of probability that a data point occurs in a cluster is used to define the distance between cluster center and the data point. Chatzis [24] propose a FCM type clustering algorithm for mixed data that employs a probabilistic dissimilarity function in a FCM-type fuzzy clustering cost function proposed by Honda and Ichihashi [68].

Kuri-Moraleset al. [102] propose a strategy for the assignment of numerical value to a categorical value. Firstly, a mixed dataset is converted into a pure numeric dataset then fuzzy C-means clustering algorithm is used.

Pathak and Pal [143] combine fuzzy, probabilistic and collaborative clustering framework for clustering mixed data. Fuzzy clustering is used to cluster numeric data portion of the mixed data, whereas mixture models [13], [24] are used to cluster categorical data portion of the mixed data. Collaborative clustering [144] is used to find the common cluster sub-structures in the categorical and numerical data.

6) Other developments

In this subsection, we discuss developments that do not fit well in our classification.

Constraint-based clustering [171] group similar data points into several clusters under certain user constraints such as two given data points will be a part of the same cluster. Cheng and Leu [31] propose a constrained K-prototypes clustering algorithm that simultaneously handles user constraints and mixed data. The algorithm extends K-prototypes clustering algorithm [78] by adding a constrained function to the cost function of the K-prototypes.

Ciaccio et al. [35] extend the well-separated partition definition [88] to propose a non-hierarchical clustering algorithm for mixed data. Sowjanya and Shashi [165] propose an incremental clustering approach for mixed data. Initially, some data points are clustered and other data points are assigned to clusters depending upon their distances from the cluster centers that are updated as new data points join the clusters. A cluster center is defined, for a categorical feature, by using mode of categorical values of data points present in the cluster. For a numeric feature, the mean of the values of data points present in a cluster is used to represent the center of the cluster. It is not clear in the paper which distance measure is used to cluster data points.

Frey and Dueck [54] propose affinity propagation clustering (APC) algorithm that uses message passing. Zhang and Gu [180] extend this method by combining the distance measure proposed by Ahmad and Dey [3] with APC algorithm. Accurate clustering results are achieved with this method. He at al. [67] extend Squeezer algorithm [65] which works for pure categorical datasets for clustering mixed data. In one of the versions, the numeric features are discretized to convert them into categorical features and then Squeezer algorithm is applied on the new categorical data. In another work, He et al. [66] divide the mixed data into two parts: pure numeric features and pure categorical features. Graph partitioning algorithm is used to cluster numeric data, whereas categorical data is clustered by using Squeezer algorithm. The clustering results are combined and treated as categorical data, which is clustered by using Squeezer algorithm to get the final clustering results. Hai and Susumu [60] parallelize the clustering algorithm proposed by He at al. [67] to handle large datasets.

Bohm et al. propose [15] a parameter-free clustering algorithm, INTEGRATE, for mixed data. The algorithm is based on a concept of information theory, the MDL [155]. This allows the balancing of the effect of both kinds of attributes (numeric and categorical). INTEGRATE is scalable to large datasets. Plant [146] propose a clustering algorithm Scenic (Scale-free Dependency Clustering) for mixed data. Mixed-type feature dependency patterns are detected by projecting the data points and the features into a joint low-dimensional feature space [130]. Then, the clusters are searched in new low-dimensional embedding.

Li and Ye [108] propose an incremental clustering approach for mixed data. Two different distance measures are proposed to compute the distance between clusters. In the first distance measure, separate distance measures are computed for numeric and categorical features, and then they are integrated into a new distance measure. In the second distance measure, categorical features are transformed into numeric features, and then a distance measure is computed by using all features. Similar clustering results are achieved with both the distance measures. Mohanavalli and Jaisakthiusing [133] use Chi-square statistics for computing the weight of each feature of mixed data. The Euclidean distance for numeric features and Hamming distance for categorical features along with these weights are used to compute the distances. The authors did not write about the clustering algorithm used in their paper.

Cheung and Jia [32] present a general clustering framework that uses the concept of data point-cluster similarity and propose a unified similarity metric for mixed datasets. Accordingly, an iterative clustering algorithm is proposed that finds the number of clusters automatically. Sangam and Om [161] present a sampling based clustering algorithm for mixed datasets. The algorithm has two steps; in the first step, a sample of data points is used for clustering. In the second step, other points are assigned to the clusters depending upon their similarity with the clusters. They develop a hybrid similarity measure to determine the similarity between a data point and a cluster. The clustering algorithm presented by Cheung and Jia [32] is used in the first step.

Lin et al. [112] presents a tree-ensembles clustering algorithm, CRAFTER, for clustering high dimensional mixed datasets. In the first step, a random subset of data points is drawn and random forests clustering algorithm [163] is applied. The clustered data points are used to train tree classifiers. These trained tree-ensembles are used to cluster all the data points.

III. MOST CITED CLUSTERING ALGORITHMS, DATASETS AND PERFORMANCE METRICS

In this section, we discuss some of the most cited mixed data clustering algorithms and the common mixed datasets used in majority of the literature. We also discuss and compare the performance of many clustering algorithms on different mixed datasets.

A. MOST CITED ALGORITHMS

In Section II, we discussed different types of mixed-data clustering algorithms. As per our research findings, some types of algorithms are more cited than others. We set the Google Scholar citation as the criteria for mixed-data clustering algorithm reviewed in this paper. The papers with at least 100 citations (11^{th} November 2018) are chosen and their relevant information is shown in Table 7. Out of seven selected clustering algorithms, four are K-means type clustering algorithm, two are hierarchical clustering algorithms whereas model based clustering and neural networks-based clustering groups have one algorithm each. This indicates that K-means clustering type algorithms are more cited for mixed datasets. The low computational complexity of these algorithms could be one of the reasons for their widespread citation.

B. PERFORMANCE MEASURES

In an ideal clustering scenario, clustering labels are not available. In that situation, it is very difficult to evaluate the performance of a given clustering algorithm. Typically, the datasets that have been used to show mixed-data clustering results have class labels. These class labels are not used to perform clustering but are treated as ground truth. The final clustering results are matched with the ground truth to evaluate the performance of clustering algorithms. Many performance measures have been used in the mixed data clustering literature; F-Measure, Normalized Mutual Information, Rand

TABLE 7. Most cited clustering algorithms for mixed datasets

Number	Title of the paper	Type of clustering algorithm	Number of citations on (11^{th} November 2018)
1	"Extensions to the K-Means Algorithm for Clustering Large Data Sets with Categorical Values" [78]	K-means clustering type algorithm	2095
1	"Bayesian classification (AutoClass): theory and results" [26]	Model based clustering	1775
3	"Automated variable weighting in K-means type clustering" [76]	K-means clustering type algorithm	680
4	"A Robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment" [49]	Hierarchical clustering algorithm	621
5	A "K-means clustering algorithm for mixed numeric and categorical data" [3]	K-means clustering type algorithm	420
6	"Feature Weighting in K-Means Clustering" [133]	K-means clustering type algorithm	361
7	"Unsupervised learning with mixed numeric and nominal data" [106]	Hierarchical clustering algorithm	190
8	"Generalizing Self-Organizing Map for Categorical Data" [70]	Neural networks-based clustering algorithm	122

TABLE 8. Mixed datasets used in the comparative study of various mixed data clustering algorithms.

Dataset name	Number of data points	Number of (categorical features)	Number of (numeric features)	Classes
Heart (Statlog project)	270	7	6	2
Heart (Cleveland)	303	8	5	2
Australian Credit	690	8	6	2
Adult	45222	8	6	2

index, etc. [87]. However, clustering accuracy has been most commonly used criterion to evaluate the quality of clustering results. The clustering accuracy (AC) is calculated by using the following formula;

$$AC = \sum_{i=1}^{n} c_i/n \tag{2}$$

where c_i is the number of data points occurring both in i^{th} cluster and its corresponding true class, and n is the number of data points in the dataset. The assignment of a class label to a cluster is done such that the AC is maximum. In the next section, we will compare the accuracy of various clustering algorithm by using AC measure.

C. ACCURACY BASED COMPARATIVE STUDY OF CLUSTERING ALGORITHMS

It is difficult to present a comparative study among different mixed data clustering algorithms as they test their clustering algorithms on different datasets. Heart (Cleveland), Heart (Statlog) , Australian Credit data and Adult data are used in many papers. The information about these datasets is presented in Table 8. We present the comparative study of several clustering algorithms on these four datasets in Table 9. The result of each algorithm is taken from its original paper. Only those results are presented in which the used performance measure was AC. Table 9 suggests that there is no algorithm that has a clear superiority over others on all the four datasets. For example, Wei et al. [174] shows the best result for Statlog Heart dataset, whereas for Cleaveland Heart dataset, the algorithms by Ahmad and Dey [3] and Rico and Diez [9] perform best. The algorithm by Wei et al. [174] perform best for Australian credit dataset, whereas Modha and Spangler's algorithm [133] perform best for Adult dataset.

IV. SOFTWARE AND APPLICATIONS

In this section, we discuss the publicly available software tools for clustering mixed datasets. We also highlight major applications areas of mixed data clustering in this section.

Many tools are available for clustering mixed datasets. Most of these tools are in R [168]. K-prototypes clustering algorithm [78] is available in R [166]. ClustMD package in R [128] is the implementation of model based clustering for mixed Data [36]. Gower's similarity matrix [59] is implemented in R. The similarity matrix can be used with partitioning around medoids tools in R or Hierarchical clustering tools to get final clusters [149]. ClustOfVar [25] is an R package for the clustering that can handle mixed datasets. Both hierarchical clustering algorithm and a K-means type partitioning algorithm are implemented in the package. CluMix is another package in R for clustering and visualization of mixed data [80]. Implementation of KAMILA [52] clustering algorithm is available in R [53]. The mixed data clustering algorithm by Macbar et al. [123] is implemented in R [122]. Ahmad and Dey algorithm [3] is available in Matlab [6]. A K-means type clustering algorithm that can deal with mixed datasets is implemented in Matlab by using feature discretization [14]. MixtComp is C++ implementation of Model-based cluster-

TABLE 9. Results of mixed clustering algorithms for various datasets. "-" shows that the result for this dataset is not present in the related paper. $\leq$ results suggest that the algorithm is run many times and all results are less than the given value. Bold number shows the best performance for that dataset.

Algorithm	**Heart (Statlog)**	**Heart (Cleveland)**	**Australian Credit**	**Adult**
Huang [78]	-	-	$>$.70	-
Amir and Dey [3]	-	**0.85 (average)**	0.88 (average)	-
Modha and Spangler [133]	-	0.83 (best case)	0.83 (best case)	**0.76**
Huang et al. [76]	$\leq$0.85	-	$\leq$0.85	-
Lam et al. [104]	-	0.80	0.79	-
Jia and Cheung [92]	0.84	-	0.85	0.75
SBAC [106]	-	0.75	-	-
Wang et al. [172]	0.87 (average)	-	0.81 (average)	-
Du et al. [44]	0.82	-	0.86	-
Foss et al. [52]	-	-	0.80	-
Wei et al. [174]	**0.89 (average)**	0.65 (average)	**0.91 (average)**	-
Chatzis [24]	-	0.81 (average)	0.85 (average)	-
He et al. [66]	-	0.83	0.77	-
He et al. [67] (dsqueezer)	-	0.83	0.57	-
Ahmad and Hashmi [64]	0.81	0.84	.086	-
Pathak and Pal [143]	0.82	0.79	0.88	-
Rico and Diez [9]	-	**0.85**	0.87	-
Ji et al. [91]	-	0.83	0.83	-

ing of mixed data [12].

A. MAJOR APPLICATION AREAS

Mixed datasets are available in different application domains, such as health, marketing, business, finance, social studies, etc. Researchers have applied various mixed data clustering algorithms on these datasets. Below, we present a list of major application areas where mixed-data clustering is mostly applied.

a: Health and Biology

McParland et al [36], [126] develop mixed data clustering algorithm to study high dimensional numeric phenotypic data and categorical genotypic data. The study leads to a better understanding of metabolic syndrome (MetS). Malo et al. [121] use mixed data clustering to study people who died of cancer between 1994 and 2006 in Hijuelas. Saadaoui et al. [159] develop a mixed data clustering algorithm for heterogeneous occupational medicine data mining. Researchers have used various types of clustering approaches for mixed data for heart disease [3], [4], [106], [133], Occupational Medicine [158], digital mammograms [61], Acute Inflammations [91], [143], [146], age of abalone snails [146], Human life span [103], Dermatology [147], medical diagnosis [108], Toxicogenomics [19], genetic regulation, analysis of biomedical datasets, [7], cancer samples grouping [179], etc.

b: Business and Marketing

Hennig and Liao [34] apply mixed data clustering techniques for socio-economic stratification by using 2007 US Survey data of consumer finances. Kassi et al. [97] develop mixed data clustering algorithm to segment gasoline services stations in Morocco to determine important features that can influence the profit of these service stations. Mixed data clustering has also been used in Credit Approval [3], [4], [76], [106], [133], Income prediction (Adult data) [75], [89], [133], Marketing Research [134], Customer Behavior Discovery [30], customer segmentation and catalog marketing [71], customer behavior pattern discovering [29], motor Insurance [79] and construction management [31].

c: Others

Moustaki and Papageorgiou [135] apply mixed data clustering in Archaeometry for classifying archaeological findings into groups. Philip and Ottaway [145] use mixed data clustering to cluster cypriot hooked-tang weapons. Chiodi use mixed data clustering for andrological data [33]. Iam-On and Boongoen [85] use mixed data clustering for student dropout prediction in a Thai university. Mixed data clustering has also been used in teaching assistant evaluation [104], [110], class examination [83], petroleum recovery [104], intrusion detection [108], [113], [153], forest cover type [95], online learning systems [139], automobiles [147], printing process delays [9], country flags mining [107], etc.

V. CHALLENGES, FUTURE DIRECTIONS AND SUMMARY

In the previous sections, we mentioned several technical challenges for mixed-data clustering algorithms, in terms of their time complexity, repeatability of results, knowledge of number of clusters and publicly available datasets for experiments. Below, we summarize our findings based on literature review and provide detailed commentary on these challenges.

- K-means type clustering algorithms have been very popular for mixed-data clustering due to their linear time complexity. However, the definition of a cluster center that serve as a good representative of a cluster needs research [3], [78].
- Various distance measures have been developed to compute the distance measure between a cluster center and a data point for K-means clustering type algorithms for mixed datasets [3], [76], [78]. The development of better distance measures that can accurately capture the

distance between a cluster center and a data point is an important future research direction.

- The distance measures that directly combine numeric and categorical distances needs more understanding in terms of the information and their scales that are being joined together.
- The random initialization of cluster centers is a problem for K-means clustering type algorithms. Few methods have been proposed to address this problem for K-means clustering type algorithms for mixed datasets [28], [64], [90]; however, either these methods are computationally expensive or do not give consistent results in different runs. The development of new methods that give fixed initial centers and good clustering performance is required.
- The number of clusters is a user defined parameter for K-means clustering type algorithms for mixed datasets. Few methods have been proposed to address this problem [110], [117]. Research effort is needed to determine the natural number of clusters in the mixed-dataset scenario.
- Getting an accurate similarity matrix is important for hierarchical clustering. Various similarity measures have been proposed for this purpose [49], [58], [59]. However, detailed experimental and theoretical studies are required to understand these similarity measures. Development of new accurate similarity measures is an important research field in mixed data clustering.
- Few subspace clustering methods have been developed for clustering high dimensional mixed datasets [4], [92]. The development of new methods which can produce more accurate clustering results is a promising research direction.
- Cluster ensembles have been used to combine the results of several clustering algorithms [57], [170]. The final clustering result is generally better than individual clustering result. The application of clustering ensembles for mixed data clustering is a new and interesting research area.
- Many mixed datasets have missing values. Development of clustering algorithms to handle missing data is a very relevant research direction [82].
- Many model based mixed data clustering are proposed that suffer from large complexity [26], [48]. Developing model based mixed data clustering with lower complexity is important to use them to solve real world problems.
- Generally, few datasets are used for the comparative study of different mixed data clustering algorithms. Therefore, efforts are needed to create mixed data repositories to help in evaluating and comparing different mixed-data clustering algorithms.
- The implementations of most of mixed data clustering are not publicly available. Therefore, the comparative study of different mixed data clustering algorithms is not an easy task. Providing relevant code on publicly available repositories will help in making fast advancement in this field.
- Recently, few papers [104], [174] have suggested transforming a mixed dataset into a numeric dataset so that clustering algorithms for numeric datasets can be used. The development of effective transformation techniques without loss of information is an important step for the success of these types of algorithms.
- Only a few clustering algorithms discuss the problem of interpretability of clustering results for mixed datasets; to describe why a certain set of data points form a cluster and how different clusters can be distinguished from each other effectively [148]. Development of new clustering algorithms that can facilitate interpretation of the clustering results is an interesting research area.
- The development of scalable mixed data clustering algorithms is key to handle present day's challenges of big data analysis. Parallelization of mixed data clustering algorithm [95] is an important research area.
- A few papers have discussed mixed data clustering algorithms for mixed data streams [27]. This is another important research area to perform mixed data clustering in an online manner.
- Few spectral clustering algorithms and constrained clustering algorithms for mixed datasets have been developed. However, the development of these types of mixed datasets is an important research area.
- Not all features of a mixed dataset may be important, removing insignificant features can improve the clustering results. Unsupervised feature selection for mixed datasets is not explored much and can be an important research area to deal with datasets having a large number of features.

VI. CONCLUSION

Mixed datasets occur more frequently than thought in several real world applications. Clustering these datasets is often desired to discover groups and their relationships. However, clustering mixed datasets is a challenging task because of the presence of both numeric and categorical features in these datasets. In this paper, we present a taxonomy to categorize different mixed data clustering algorithms based on the methodology adopted to create clusters. Then, we presented a comprehensive review of these algorithms. Experimental comparison of these algorithm was also presented in the paper. Publicly available tools for mixed data clustering were also discussed in the paper to aide researchers test some of the standard clustering methods on their datasets. Future directions of this area were also presented in detail, which will be helpful for researchers in this area. Through this survey, we highlighted different type of mixed-data clustering algorithms, their advantages and shortcomings, and presented many plausible research ideas to make progress in this field. This survey should be able to guide researchers to develop an in-depth understanding of the field and generate new ideas to make significant contributions to solve real world problems.

REFERENCES

[1] C. C. Aggarwal and C. K. Reddy. Data Clustering: Algorithms and Applications. Chapman and Hall/CRC, 2013.

[2] A. Ahmad and L. Dey. Algorithm for Fuzzy Clustering of Mixed Data with Numeric and Categorical Attributes, pages 561–572. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.

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

[4] A. Ahmad and L. Dey. A k-means type clustering algorithm for subspace clustering of mixed numeric and categorical datasets. Pattern Recognition Letters, 32(7):1062–1069, 2011.

[5] D. Alahakoon, S. K. Halgamuge, and B. Srinivasan. Dynamic self-organizing maps with controlled growth for knowledge discovery. Trans. Neur. Netw., 11(3):601–614, May 2000.

[6] Ahmad Alsahaf. mixed kmeans package. https://www.mathworks.com/matlabcentral/fileexchange/53489-amjams-mixed-kmeans?requestedDomain=www.mathworks.com, 2016. Online accessed 28-January-2018.

[7] B. Andreopoulos, A. An, and X. Wang. Bi-level clustering of mixed categorical and numerical biomedical data. IJDMB, 1(1):19–56, 2006.

[8] K. Balaji and K. Lavanya. Clustering algorithms for mixed datasets: A review. International Journal of Pure and Applied Mathematics, 18(7):547–556, 2018.

[9] F. Barcelo-Rico and D. Jose-Luis. Geometrical codification for clustering mixed categorical and numerical databases. Journal of Intelligent Information Systems, 39(1):167–185, 2012.

[10] J. C. Bezdek. Pattern Recognition with Fuzzy Objective Function Algorithms, chapter Pattern Recognition with Fuzzy Objective Function. 1981.

[11] C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(7):719–725, Jul 2000.

[12] C. Biernacki and V. Kubicki. Mixtcomp software for full mixed data. https://modal.lille.inria.fr/wikimodal/doku.php?id=mixtcomp, 2016. C++ package- Online accessed 28-January-2018.

[13] C. M. Bishop. Pattern Recognition and Machine Learning. Springer-Verlag New York Inc, 2008.

[14] Camden Bock. Mixed k-means clustering algorithm with variable discretization. https://www.mathworks.com/matlabcentral/fileexchange/55601-mixed-k-means-clustering-algorithm-with-variable-discretization, 2016. Matlab package- Online accessed 28-January-2018.

[15] C. Böhm, S. Goebl, A. Oswald, C. Plant, M. Plavinski, and B. Wackersreuther. Integrative Parameter-Free Clustering of Data with Mixed Type Attributes, pages 38–47. 2010.

[16] 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, pages 243–254.

[17] M. Boris. Reinterpreting the category utility function. Machine Learning, 45(2):219–228, 2001.

[18] R. P. Browne and P. D. McNicholas. Model-based clustering, classification, and discriminant analysis of data with mixed type. Journal of Statistical Planning and Inference, 142(11):2976–2984, 2012.

[19] P. R. Bushel. Clustering of Mixed Data Types with Application to Toxicogenomics. PhD thesis, North Carolina State University, 2006.

[20] B.Zhang. Generalized <italic>K</italic>-Harmonic Means, Dynamic Weighting of Data in Unsupervised Learning, pages 1–13. 2001.

[21] T. Calinski and J. Harabasz. A dendrite method for cluster analysis. Communications in Statistics, 3(1):1–27, 1974.

[22] G. A. Carpenter and S. Grossberg. Adaptive resonance theory. In Encyclopedia of Machine Learning, pages 22–35. 2010.

[23] 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 Networks, 4(6):759 – 771, 1991.

[24] S. P. Chatzis. A fuzzy c-means-type algorithm for clustering of data with mixed numeric and categorical attributes employing a probabilistic dissimilarity functional. Expert Systems with Applications, 38(7):8684–8689, 2011.

[25] M. Chavent and J. Saracco V. K. Simonet, B. Liquet. Clustofvar: An r package for the clustering of variables. Journal of Statistical Software, 50(13):1–16, 2012.

[26] P. Cheeseman and J. Stutz. Advances in Knowledge Discovery and Data Mining, chapter Bayesian Classification (AutoClass): Theory and Results, pages 153–180. 1996.

[27] J. Chen and H. He. A fast density-based data stream clustering algorithm with cluster centers self-determined for mixed data. Information Sciences, 345:271Ű293, 2016.

[28] J. Chen, X. Lin Xiang, H. Zheng, and X. Bao. A novel cluster center fast determination clustering algorithm. Applied Soft Computing, 57:539–555, 2017.

[29] M. Cheng, Y. Xin, Y. Tian, C. Wang, and Y. Yang. Customer behavior pattern discovering based on mixed data clustering. In 2009 International Conference on Computational Intelligence and Software Engineering, pages 1–4, Dec 2009.

[30] Mingzhi Cheng, Yang Xin, Yangge Tian, Cong Wang, and Yixian Yang. Customer behavior pattern discovering based on mixed data clustering. In Computational Intelligence and Software Engineering, 2009. CiSE 2009. International Conference on, pages 1–4. IEEE, 2009.

[31] Y. Cheng and S. Leu. Constraint-based clustering and its applications in construction management. Expert Systems with Applications, 36(3, Part 2):5761 – 5767, 2009.

[32] Y. Cheung and H. Jia. Categorical-and-numerical-attribute data clustering based on a unified similarity metric without knowing cluster number. Pattern Recognition, 46(8):2228 – 2238, 2013.

[33] M. Chiodi. A partition type method for clustering mixed data. Rivista di Statistica Applicata, 2:135Ű147, 1990.

[34] C. H. Christian and T. F. Liao. How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. Journal of the Royal Statistical Society Series C, 62(3):309–369, 2013.

[35] A. D. Ciaccio. MIXISO: a Non-Hierarchical Clustering Method for Mixed-Mode Data, pages 27–34. Springer Berlin Heidelberg, Berlin, Heidelberg, 2001.

[36] C.M. Phillips D. McParland, L. Brennan, H.M. Roche, and I. C. Gormley. Clustering high dimensional mixed data to uncover sub-phenotypes: joint analysis of phenotypic and genotypic data. CoRR, abs/1604.01686, 2016.

[37] G. David and A. Averbuch. Spectralcat: Categorical spectral clustering of numerical and nominal data. Pattern Recognition, 45(1):416 – 433, 2012.

[38] J. Dean and S. Ghemawat. Mapreduce: Simplified data processing on large clusters. Communications of the ACM, 51(1):107–113, 2008.

[39] Carmelo del Coso, Diego Fustes, Carlos Dafonte, Francisco J. Nóvoa, José M. Rodríguez-Pedreira, and Bernardino Arcay. Mixing numerical and categorical data in a self-organizing map by means of frequency neurons. Applied Soft Computing, 36:246 – 254, 2015.

[40] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B, 39(1):1–38, 1977.

[41] H. P. Devaraj and M. Punithavalli. An integrated framework for mixed data clustering using self organizing map. Journal of Computer Science, 7(11):1639–1645, 2011.

[42] S. Ding, M. Du, T. Sun, X. Xu, and Y. Xue. An entropy-based density peaks clustering algorithm for mixed type data employing fuzzy neighborhood. Knowledge-Based Systems, 133:294 – 313, 2017.

[43] C. Doring, C. Borgelt, and R. Kruse. Fuzzy clustering of quantitative and qualitative data. In Fuzzy Information, 2004. Processing NAFIPS '04. IEEE Annual Meeting of the, volume 1, pages 84–89 Vol.1, 2004.

[44] M. Du, S. Ding, and Y. Xue. A novel density peaks clustering algorithm for mixed data. Pattern Recognition Letters, 97:46 – 53, 2017.

[45] J. C. Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. Journal of Cybernetics, 3(3):32–57, 1973.

[46] Y. El-Sonbaty and M. A. Ismail. Fuzzy clustering for symbolic data. IEEE Transactions on Fuzzy Systems, 6(2):195–204, 1998.

[47] M. Ester, H. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. pages 226–231. AAAI Press, 1996.

[48] B. S. Everitt. A finite mixture model for the clustering of mixed-mode data. Statistics and Probability Letters, 6(5):305–309, 1988.

[49] T. Chiu D. P. Fang, J. Chen, Y. Wang, and C. Jeris. A robust and scalable clustering algorithm for mixed type attributes in large database environment. In Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '01, pages 263–268, 2001.

[50] M. Filippone, F. Camastra, F. Masulli, and S. Rovetta. A survey of kernel and spectral methods for clustering. Pattern Recognition, 41(1):176 – 190, 2008.

[51] D. H. Fisher. Knowledge acquisition via incremental conceptual clustering. Machine Learning, 2(2):139–172, Sep 1987.

[52] A. Foss, M. Markatou, B. Ray, and A. Heching. A semiparametric method for clustering mixed data. Machine Learning, 105(3):419–458, 2016.

[53] Alexander Foss and Marianthi Markatou. kamila: Methods for clustering mixed-type data. https://cran.r-project.org/web/packages/kamila/index.html, 2016. R package- Online accessed 28-January-2018.

[54] Brendan J. Frey and Delbert Dueck. Clustering by passing messages between data points. Science, 315:2007, 2007.

[55] S. Furao and O. Hasegawa. An incremental network for on-line unsupervised classification and topology learning. Neural Networks, 19(1):90–106, 2006.

[56] J. H. Gennari, P. Langley, and D. Fisher. Models of incremental concept formation. Artificial Intelligence, 40(1):11–61, 1989.

[57] J. Ghosh and A. Acharya. Cluster ensembles. Wiley Interdisc. Rew.: Data Mining and Knowledge Discovery, 1(4):305–315, 2011.

[58] D.W. Goodall. A new similarity index based on probability. Biometrics, 22:882–907, 1966.

[59] J. C. Gower. A general coefficient of similarity and some of its properties. Biometrics, 27(4):857–871, 1971.

[60] N. T. M. Hai and H. Susumu. Performances of Parallel Clustering Algorithm for Categorical and Mixed Data, pages 252–256. 2005.

[61] S.M. Halawani, M. Alhaddad, and A. Ahmad. A study of digital mammograms by using clustering algorithms. Journal of Scientific and Industrial Research (JSIR), 71:594–600, 2012.

[62] J. Han, Y. Cai, and N. Cercone. Data-driven discovery of quantitative rules in relational databases. IEEE Transactions on Knowledge and Data Engineering, 5(1):29–40, 1993.

[63] J. Han and Y. Fu. Dynamic generation and refinement of concept hierarchies for knowledge discovery in databases. In in: Proceedings of the AAAIŠ94 Workshop Knowledge Discovery in Databases (KDDŠ94), page 157Ű168, 1994.

[64] A. Ahmad S. Hashmi. K-harmonic means type clustering algorithm for mixed datasets. Applied Soft Computing, 48(C):39–49, 2016.

[65] Z. He, X. Xu, and S. Deng. Squeezer: An efficient algorithm for clustering categorical data. Journal of Computer Science and Technology, 17(5):611–624, 2002.

[66] Z. He, X. Xu, and S. Deng. Clustering Mixed Numeric and Categorical Data: A Cluster Ensemble Approach. eprint arXiv:cs/0509011, 2005.

[67] Z. He, X. Xu, and S. Deng. Scalable algorithms for clustering large datasets with mixed type attributes. International Journal of Intelligent Systems, 20(10):1077–1089, 2005.

[68] K. Honda and H. Ichihashi. Regularized linear fuzzy clustering and probabilistic pca mixture models. IEEE Transactions on Fuzzy Systems, 13(4):508–516, 2005.

[69] E. R. Hruschka, R. J. G. B. Campello, A. A. Freitas, and A. C. Ponce Leon F. de Carvalho. A survey of evolutionary algorithms for clustering. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 39(2):133–155, March 2009.

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

[71] C. Hsu and Y. Chen. Mining of mixed data with application to catalog marketing. Expert Systems with Applications, 32(1):12 – 23, 2007.

[72] C. Hsu and S. Lin. Visualized analysis of multivariate mixed-type data via an extended self-organizing map. In The 6th International Conference on Information Technology and Applications (ICITA 2009), pages 218–223, 2006.

[73] C. Hsu and S. Lin. Visualized analysis of mixed numeric and categorical data via extended self-organizing map. IEEE Transactions on Neural Networks and Learning Systems, 23(1):72–86, 2012.

[74] C. C. Hsu, C. G. Chen, and Y. Su. Hierarchical clustering of mixed data based on distance hierarchy. Information Sciences, 177(20):4474–4492, 2007.

[75] C. C. Hsu and Y. P. Huang. Incremental clustering of mixed data based on distance hierarchy. Expert Systems with Applications, 35(3):1177 – 1185, 2008.

[76] J. Z. Huang, M. K. Ng, Hongqiang Rong, and Zichen Li. Automated variable weighting in k-means type clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(5):657–668, 2005.

[77] Z. Huang. Clustering large data sets with mixed numeric and categorical values. In Proceedings of the First Pacific Asia Knowledge Discovery and Data Mining Conference, pages 21–34. Singapore: World Scientific, 1997.

[78] Z. Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Min. Knowl. Discov., 2(3):283–304, 1998.

[79] Zhexue Huang. Clustering large data sets with mixed numeric and categorical values. In In The First Pacific-Asia Conference on Knowledge Discovery and Data Mining, 1997.

[80] M. Hummel, D. Edelmann, and A. Kopp-Schneider. Clumix: Clustering and visualization of mixed-type data. https://cran.r-project.org/web/packages/CluMix/index.html, 2017. R package- Online accessed 28-January-2018.

[81] L. Hunt and M. Jorgensen. Mixture model clustering of data sets with categorical and continuous variables. In Information, Statistics and Induction in Science, page 375Ű384. Singapore: World Scientific, 1996.

[82] L. Hunt and M. Jorgensen. Mixture model clustering for mixed data with missing information. Computational Statistics and Data Analysis, 41(3âĂŞ4):429–440, 2003.

[83] L. Hunt and M. Jorgensen. Clustering mixed data. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(4):352–361, 2011.

[84] Isabella I. Morlini and S. Zani. Comparing Approaches for Clustering Mixed Mode Data: An Application in Marketing Research, pages 49–57. 2010.

[85] N. Iam-On and T. Boongoen. Improved student dropout prediction in thai university using ensemble of mixed-type data clusterings. International Journal of Machine Learning and Cybernetics, 8(2):497–510, 2017.

[86] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. ACM Comput. Surv., 31(3):264–323, September 1999.

[87] A.K. Jain and R.C. Dubes. Algorithms for Clustering Data. Prentice Hall, 1988.

[88] N. Jardine and R. Sibson. Mathematical Taxonomy. Wiley London, 1971.

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

[90] J. Ji, W. Pang, Y. Zheng, Z. Wang, Z. Ma, and L. Zhang. A novel cluster center initialization method for the k-prototypes algorithms using centrality and distance. Applied Mathematics and Information Sciences, 9(6):2933–2942, 2015.

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

[92] H. Jia and Y. M. Cheung. Subspace clustering of categorical and numerical data with an unknown number of clusters. IEEE Transactions on Neural Networks and Learning Systems, PP(99):1–18, 2017.

[93] L. Jing, M. K. Ng, and J. Z. Huang. An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data. IEEE Trans. on Knowl. and Data Eng., 19(8):1026–1041, 2007.

[94] S. San K J. Chae and W. Y. Yang. Cluster analysis with balancing weight on mixed-type data. The Korean Communications in Statistics, 13(3):719–732, 2006.

[95] M. A. B. Kacem, C. E. B. N'cir, and N. Essoussi. Mapreduce-based k-prototypes clustering method for big data. In 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pages 1–7, 2015.

[96] A. H. Kashan. League championship algorithm: A new algorithm for numerical function optimization. In 2009 International Conference of Soft Computing and Pattern Recognition, pages 43–48, 2009.

[97] M. L. Kassi, A. Berrado, L. Benabbou, and K. Benabdelkader. Towards a new framework for clustering in a mixed data space: Case of gasoline service stations segmentation in morocco. In 2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA), pages 1–6, 2015.

[98] S. S. Khan and A. Ahmad. Cluster center initialization algorithm for k-modes clustering. Expert Syst. Appl., 40(18):7444–7456, 2013.

[99] Shehroz S. Khan and Amir Ahmad. Cluster center initialization algorithm for k-means clustering. Pattern Recognition Letters, 25(11):1293–1302, 2004.

[100] T. Kohonen. Self-organized formation of topologically correct feature maps. Biological Cybernetics, 43(1):59–69, 1982.

[101] T. Kohonen, M. R. Schroeder, and T. S. Huang, editors. Self-Organizing Maps. Springer-Verlag, Berlin, Heidelberg, 3rd edition, 2001.

[102] Angel Kuri-Morales, Daniel Trejo-Baños, and Luis Enrique Cortes-Berrueco. Clustering of heterogeneously typed data with soft computing - a case study. In Proceedings of the 10th International Conference on Artificial Intelligence: Advances in Soft Computing - Volume Part II, pages 235–248. Springer-Verlag, 2011.

[103] Angel Kuri-Morales, Daniel Trejo-Baños, and Luis Enrique Cortes-Berrueco. Clustering of heterogeneously typed data with soft computing - a case study. In Proceedings of the 10th International Conference on Artificial Intelligence: Advances in Soft Computing - Volume Part II, MICAI'11, pages 235–248, Berlin, Heidelberg, 2011. Springer-Verlag.

[104] D. Lam, M. Wei, and D. Wunsch. Clustering data of mixed categorical and numerical type with unsupervised feature learning. IEEE Access, 3:1605–1613, 2015.

[105] C. J. Lawrence and W. J. Krzanowski. Mixture separation for mixed-mode data. Statistics and Computing, 6(1):85–92, 1996.

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

[107] Taoying Li and Yan Chen. A weight entropy k-means algorithm for clustering dataset with mixed numeric and categorical data. In Fuzzy Systems and Knowledge Discovery, 2008. FSKD'08. Fifth International Conference on, volume 1, pages 36–41. IEEE, 2008.

[108] X. Li and N. Ye. A supervised clustering and classification algorithm for mining data with mixed variables. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 36(2):396–406, 2006.

[109] J. Liang, K. S. Chin, C. Dang, and R. C. M. Yam. A new method for measuring uncertainty and fuzziness in rough set theory. International Journal of General Systems, 31(4):331–342, 2002.

[110] J. Liang, X. Zhao, D. Li, F. Cao, and C. Dang. Determining the number of clusters using information entropy for mixed data. Pattern Recognition, 45(6):2251–2265, 2012.

[111] J. Lim, J. Jun, S.H. Kim, and D. McLeod. A framework for clustering mixed attribute type datasets. In Proc. of the 4th Int. Con. on Emerging Databases (EDB 2012), 2012.

[112] S. Lin, B. Azarnoush, and G. Runger. Crafter: a tree-ensemble clustering algorithm for static datasets with mixed attributes and high dimensionality. IEEE Transactions on Knowledge and Data Engineering, Article in Press.

[113] N. Liu. The Research of Intrusion Detection Based on Mixed Clustering Algorithm, pages 92–100. 2012.

[114] X. Liu, Q. Yang, and L. He. A novel dbscan with entropy and probability for mixed data. Cluster Computing, 20(2):1313–1323, Jun 2017.

[115] Y. Lu, S. Lu, F. Fotouhi, Y. Deng, and J. S. Brown. Fgka: A fast genetic k-means clustering algorithm. In Proceedings of the 2004 ACM Symposium on Applied Computing, pages 622–623, New York, NY, USA, 2004. ACM.

[116] H. Luo, F. Kong, and Y. Li. Clustering Mixed Data Based on Evidence Accumulation, pages 348–355. Springer Berlin Heidelberg, 2006.

[117] J.E. Corter M.A. Gluck. Information, uncertainty, and the utility of categories. In Proceeding of the 7th Annual Conference of the Cognitive Science Society, Lawrence Erlbaum Associates, Irvine, pages 283–287, 1985.

[118] M.Z. Islam M.A. Rahman. Crudaw: a novel fuzzy technique for clustering records following user defined attribute weights. In Data Mining and Analytics 2012 (AusDM 2012), Sydney, Australia, 2012, pages 27–42, 2012.

[119] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In L. M. Le Cam and J. Neyman, editors, Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability, volume 1, pages 281–297. University of California Press, 1967.

[120] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, page 281Ű297, 1967.

[121] E. Malo, R. Salas, M. Catalán, and P. López. A mixed data clustering algorithm to identify population patterns of cancer mortality in hijuelas-chile. In Proceedings of the 11th Conference on Artificial Intelligence in Medicine, AIME '07, pages 190–194, 2007.

[122] M. Marbac, C. Biernacki, and V. Vandewalle. Copules-package: Mixed data clustering by a mixture model of gaussian copulas. https://rdrr.io/rforge/MixCluster/man/Copules-package.html, 2014. R package- Online accessed 28-January-2018.

[123] M. Marbac, C. Biernacki, and V. Vandewalle. Model-based clustering of gaussian copulas for mixed data. Communications in Statistics - Theory and Methods, 46(23):11635–11656, 2017.

[124] K. Mckusick and K. Thompson. Cobweb/3: A portable implementation. Technical Report FIA-90-6-18-2, NASA Ames Research Center, 1990.

[125] G. McLachlan and T. Krishnan. The EM Algorithm and Extensions. WILEY, 2008.

[126] D. McParland and I. C. Gormley. Model based clustering for mixed data: clustmd. Adv. Data Analysis and Classification, 10(2):155–169, 2016.

[127] D. McParland, C. M. Phillips, L. Brennan, H. M. Roche, and I. C. Gormle. Clustering high-dimensional mixed data to uncover sub-phenotypes: joint analysis of phenotypic and genotypic data. Statistics in Medicine, 36(28):4548Ű4569, 2017.

[128] Damien McParland and Isobel Claire Gormley. clustmd: Model based clustering for mixed data. https://cran.r-project.org/web/packages/clustMD/index.html, 2017. R package- Online accessed 28-January-2018.

[129] V. Melnykov and R. Maitra. Finite mixture models and model-based clustering. Statistical Survey, 4(80-116), 2010.

[130] G. Michailidis and J. de Leeuw. The gifi system of descriptive multivariate analysis. Statist. Sci., 13(4):307–336, 11 1998.

[131] M.M. Campos B.L. Milenova. Clustering large databases with numeric and nominal values using orthogonal projections. Technical report, Oracle Data Mining Technologies, Oracle Corporation, 2002.

[132] M. Mitchell. An Introduction to Genetic Algorithms (Complex Adaptive Systems). MIT Press, 1998.

[133] D. S. Modha and W. S. Spangler. Feature weighting in k-means clustering. Machine Learning, 52(3):217–237, September 2003.

[134] Isabella Morlini and Sergio Zani. Comparing approaches for clustering mixed mode data: An application in marketing research. In Data Analysis and Classification, pages 49–57. Springer, 2010.

[135] I. Moustaki and I. Papageorgiou. Latent class models for mixed variables with applications in archaeometry. Computational Statistics and Data Analysis, 48(3):659–675, 2005.

[136] Ning N. Chen and N. C. Marques. An extension of self-organizing maps to categorical data. In Progress in Artificial Intelligence, pages 304–313, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[137] R. B. Nelsen. An Introduction to Copulas (Springer Series in Statistics). Springer-Verlag, Berlin, Heidelberg, 2006.

[138] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic, NIPS'01, pages 849–856, 2001.

[139] K. Niu, Z. Niu, Y. Su, C. Wang, H. Lu, and J. Guan. A coupled user clustering algorithm based on mixed data for web-based learning systems. Mathematical Problems in Engineering, 2015, 2015.

[140] F. Noorbehbahani, S. R.l Mousavi, and A. Mirzaei. An incremental mixed data clustering method using a new distance measure. Soft Computing, 19(3):731–743, 2015.

[141] L. Parsons, E. Haque, and H. Liu. Subspace clustering for high dimensional data: A review. ACM SIGKDD Explorations Newsletter, 6(1):90–105, June 2004.

[142] Lance Parsons, Ehtesham Haque, and Huan Liu. Subspace clustering for high dimensional data: A review. SIGKDD Explor. Newsl., 6(1):90–105, June 2004.

[143] A. Pathak and N. R. Pal. Clustering of mixed data by integrating fuzzy, probabilistic, and collaborative clustering framework. International Journal of Fuzzy Systems, 18(3):339–348, 2016.

[144] W. Pedrycz. Collaborative fuzzy clustering. Pattern Recognition Letters, 23(14):1675–1686, 2002.

[145] G. Philip and B. S. Ottaway. Mixed data cluster analysis: an illustration using cypriot hooked-tang weapons. Archaeometry, 25(2):119–133, 1983.

[146] C. Plant. Dependency clustering across measurement scales. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, pages 361–369, 2012.

[147] C. Plant and C. Böhm. Inconco: Interpretable clustering of numerical and categorical objects. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '11, pages 1127–1135, 2011.

[148] Claudia Plant and Christian Böhm. Inconco: interpretable clustering of numerical and categorical objects. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 1127–1135. ACM, 2011.

[149] Wicked Good Data r. Clustering mixed data types in r. https://www.r-bloggers.com/clustering-mixed-data-types-in-r/, 2016. R package- Online accessed 28-January-2018.

[150] Md Anisur Rahman and Md Zahidul Islam. A hybrid clustering technique combining a novel genetic algorithm with k-means. Knowledge-Based Systems, 71:345–365, 2014.

[151] V. Rajan and S. Bhattacharya. Dependency clustering of mixed data with gaussian mixture copulas. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI'16, pages 1967–1973, 2016.

[152] Y. Reich and J. S. Fenves. Concept formation knowledge and experience in unsupervised learning. chapter The Formation and Use of Abstract Concepts in Design, pages 323–353. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1991.

[153] M. Ren, P. Liu, Z. Wang, and X. Pan. An improved mixed-type data based kernel clustering algorithm. In 2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), pages 1205–1209, 2016.

[154] A. Renyi. On measures of entropy and information. In Proceeding of the 4th Berkeley Symposium on Mathematics of Statistics and Probability, page 547Ű561, 1961.

[155] J. Rissanen. Modeling by shortest data description. Automatica, 14(5):465 – 471, 1978.

[156] A. Rodriguez and A. Laio. Clustering by fast search and find of density peaks. Science, 344(6191):1492–1496, 2014.

[157] D. K. Roy and L. K. Sharma. Genetic k-means clustering algorithm for mixed numeric and categorical data sets. International Journal of Artificial Intelligence, 1(2):23Ű28, 2010.

[158] Foued Saâdaoui, Pierre R Bertrand, Gil Boudet, Karine Rouffiac, Frédéric Dutheil, and Alain Chamoux. A dimensionally reduced clustering methodology for heterogeneous occupational medicine data mining. IEEE transactions on nanobioscience, 14(7):707–715, 2015.

[159] F. Saâdaoui, P. R. Bertrand, G. Boudet, K. Rouffiac, F. Dutheil, and A. Chamoux. A dimensionally reduced clustering methodology for heterogeneous occupational medicine data mining. IEEE Transactions on NanoBioscience, 14(7):707–715, 2015.

[160] V. Sandro and R. Jose. A survey of clustering ensemble algorithms. International Journal of Pattern Recognition and Artificial Intelligence, 25(03):337–372, 2011.

[161] R. S. Sangam and H. Om. Hybrid data labeling algorithm for clustering large mixed type data. Journal of Intelligent Information Systems, 45(2):273–293, 2015.

[162] G. Schwarz. Estimating the dimension of a model. The Annals of Statistics, 6(2):461–464, 03 1978.

[163] T. Shi and S. Horvath. Unsupervised learning with random forest predictors. Journal of Computational and Graphical Statistics, 15(1):118–138, 2006.

[164] M. Shih, J. Jheng, and L. Lai. A two-step method for clustering mixed categroical and numeric data. Tamkang Journal of Science and Engineering, 13(1):11–19, 2010.

[165] A M Sowjanya and M Shashi. A cluster feature-based incremental clustering approach to mixed data. Journal of Computer Science, 7(12):1875–1880, 2011.

[166] Gero Szepannek. clustmixtype: k-prototypes clustering for mixed variable-type data. https://cran.r-project.org/web/packages/clustMixType/index.html, 2017. R package- Online accessed 28-January-2018.

[167] Wei-Shen Tai and Chung-Chian Hsu. Growing self-organizing map with cross insert for mixed-type data clustering. Applied Soft Computing, 12(9):2856 – 2866, 2012.

[168] R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2008.

[169] L. S. Tekumalla, Vaibhav V. Rajan, and C. Bhattacharyya. Vine copulas for mixed data : multi-view clustering for mixed data beyond meta-gaussian dependencies. Machine Learning, 106(9):1331–1357, 2017.

[170] SANDRO VEGA-PONS and JOSÃĽ RUIZ-SHULCLOPER. A survey of clustering ensemble algorithms. International Journal of Pattern Recognition and Artificial Intelligence, 25(03):337–372, 2011.

[171] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl. Constrained k-means clustering with background knowledge. In Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, pages 577–584, 2001.

[172] C. Wang, C. Chi, W. Zhou, and R. Wong. Coupled interdependent attribute analysis on mixed data. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, AAAI'15, pages 1861–1867, 2015.

[173] T. Wangchamhan, S. Chiewchanwattana, and K. Sunat. Efficient algorithms based on the k-means and chaotic league championship algorithm for numeric, categorical, and mixed-type data clustering. Expert Systems with Applications, 90:146–167, 2017.

[174] M. Wei, T. W. S. Chow, and R. H. M. Chan. Clustering heterogeneous data with k-means by mutual information-based unsupervised feature transformation. Entropy, 17(3):1535–1548, 2015.

[175] I. H. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann San Francisco, CA, 2 edition, 2005.

[176] Rui Xu and D. Wunsch. Survey of clustering algorithms. IEEE Transactions on Neural Networks, 16(3):645–678, May 2005.

[177] M. S. Yang. A survey of fuzzy clustering. Mathematical and Computer Modelling, 18(11):1–16, 1993.

[178] H. Yin. Visom - a novel method for multivariate data projection and structure visualization. IEEE Transactions on Neural Networks, 13(1):237–243, January 2002.

[179] Z.ăAbidin, N. FatinăN., and R. D. Westhead. Flexible model-based clustering of mixed binary and continuous data: application to genetic regulation and cancer. Nucleic Acids Research, 45(7):e53, 2017.

[180] K. Zhang and X. Gu. An affinity propagation clustering algorithm for mixed numeric and categorical datasets. Mathematical Problems in Engineering, 2014.

[181] T. Zhang, R. Ramakrishnan, and M. Livny. Birch: A new data clustering algorithm and its applications. Data Mining and Knowledge Discovery, 1(2):141–182, 1997.

[182] W. Zhao, W. Dai, and C. Tang. K-centers algorithm for clustering mixed type data. In Proceedings of the 11th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, PAKDD'07, pages 1140–1147, 2007.

[183] Z. Zheng, M. Gong, J. Ma, L. Jiao, and Q. Wu. Unsupervised evolutionary clustering algorithm for mixed type data. In IEEE Congress on Evolutionary Computation, pages 1–8, 2010.

AMIR AHMAD received the PhD degree in computer science from the University of Manchester, United Kingdom. He is currently working as an assistant professor in the College of Information Technology, UAE University, Al Ain, United Arab Emirates. His research areas are machine learning, data mining, and nanotechnology.

SHEHROZ S. KHAN is working as a Scientist at Toronto Rehabilitation Institute, Canada. He earned his PhD in Computer Science with specialization in Machine Learning from the University of Waterloo, Canada. He did his Masters from National University of Ireland Galway, Republic of Ireland. Dr. Khan is also a Post-graduate Affiliate at the Vector Institute, Toronto. His main research focus is the development of machine learning and deep learning algorithms within the realms of ARIAL - Aging, Rehabilitation, and Intelligent Assisstive Living.