IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received January 27, 2019, accepted February 13, 2019, date of publication March 7, 2019, date of current version March 25, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2903568

Survey of State-of-the-Art Mixed Data Clustering Algorithms

AMIR AHMAD[1], AND SHEHROZ S. KHAN[2]
[1]College of Information Technology, United Arab Emirates University, Al Ain 15551, UAE
[2]Toronto Rehabilitation Institute, University Health Network, Toronto, ON M5G 2A2, Canada

Corresponding author: Amir Ahmad (amirahmad@uaeu.ac.ae)

This work was supported by the UAE University Start-Up Grant through Fund 31T101 under Grant G00002668.

ABSTRACT Mixed data comprises both numeric and categorical features, and mixed datasets occur frequently in many domains, such as health, finance, and marketing. Clustering is often applied to mixed datasets to find structures and to group similar objects for further analysis. However, clustering mixed data are challenging because it is difficult to directly apply mathematical operations, such as summation or averaging, to the feature values of these datasets. In this paper, we present a taxonomy for the study of mixed data clustering algorithms by identifying five major research themes. We then present the state-of-the-art review of the research works within each research theme. We analyze the strengths and weaknesses of these methods with pointers for future research directions. At last, we present an in-depth analysis of the overall challenges in this field, highlight open research questions, and discuss guidelines to make progress in the field.

INDEX TERMS Categorical features, clustering, mixed datasets, numeric features.

I. INTRODUCTION

Clustering is an unsupervised machine learning technique used to group unlabeled data into clusters that contain data points that are '*similar*' to each other and '*dissimilar*' from those in other clusters [1], [2]. Many clustering algorithms can only handle data that contain either numeric or categorical feature values [3], [4]. Numeric features can take real values, such as height, weight, and distance. Categorical features represent data that can be divided into a fixed number of categories, such as color, race, sex, profession, and blood group. Clustering algorithms group data points into clusters using some notion of '*similarity*', which can be as simple as the Euclidean distance. To compute the similarity between numeric feature values, mathematical operations (such as distances, angles, summation, or mean) are applied to them. Distance-based similarity measures are mostly used for numeric data points. Generally, categorical feature values are not inherently ordered (for example, the categorical values, red and blue). It is not possible to directly compute the distance between two categorical feature values. Therefore, computing distance-based similarity measures for categorical data is a challenging task [5]. Nevertheless, several methods have been suggested in the literature for computing the similarity of data points containing categorical features [5].

Many real-world datasets contain both numeric and categorical features; they are called *mixed datasets*. Mixed data occur frequently in many applications, such as health, marketing, medical, and finance [6]–[8]. Therefore, developing machine learning algorithms that can handle such data has become important. Clustering is a natural choice for practitioners to determine groups of mixed data points for further data analysis. However, the problem of computing the similarity of two data points becomes more difficult when the dataset contains both numerical and categorical features. An example snapshot of a typical mixed dataset is shown in Table 1. This sample 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 numeric and categorical parts and find the Euclidean distance between two data points for the numeric features and the Hamming distance for the categorical features [9]. This will enable one to find the similarity between numeric and categorical feature values, albeit separately. The next step is to combine these two measures to get one value that represents the distance between two mixed data points. However,

The associate editor coordinating the review of this manuscript and approving it for publication was Haruna Chiroma.

TABLE 1. An example mixed dataset.

Weight (kg)	Height (m)	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

combining these two types of distances directly is non-trivial, because it is not clear,

(i) whether both of the distance measures calculate a 'similar' type of similarity, or
(ii) whether the scales of these distances are similar. Therefore, the proportions in which the two distance measures are combined is non-obvious.

Hence, as the notion of similarity is not clearly defined for mixed data, performing clustering on them remains challenging.

Two major focuses of most mixed data clustering algorithms are (i) to find innovative ways to define novel measures of similarity between mixed features, and (ii) to perform clustering using existing or new techniques. Some of the earliest techniques of mixed clustering were direct extensions of partitional clustering algorithms (for example, K-means) [9], [10]. Since then, many new research themes have evolved and developed in this field of research. In this paper, we present a taxonomy to identify five broad research themes for mixed data clustering algorithms based on the methodology used to cluster mixed datasets. Using this taxonomy, we present a comprehensive review of clustering algorithms within each research theme. We present a critical analysis of the different types of mixed data clustering algorithms, and discuss their functions, strengths and weaknesses. We further identify challenges and open research questions among the different types of mixed data clustering algorithms and discuss on opportunities to make advances in the field. The main contributions of our paper are as follows:

- We identify a few other survey papers on mixed data clustering and differentiate them in our comprehensive literature review in terms of scope, taxonomy, research areas, applications, and vision for future work
- We present a new taxonomy to identify five broad research themes for the study of mixed data clustering and present a critical review of the literature on these research themes.
- We present a detailed analysis of the application areas in which mixed data clustering may have major impact.
- We present an in-depth analysis of ensuing challenges, open research questions and guidelines to be adopted to make progress in the field.

II. SURVEY OF OTHER REVIEW PAPERS

Few review articles on mixed data clustering have been published recently. However, they are not detailed and they concentrate on specific types of clustering algorithms. Velden *et al.* [11] study five distance-based clustering algorithms for mixed data on three mixed datasets. They conclude that there is no single clustering approach that performs well for all the datasets. The review presented by Foss *et al.* [12] concentrate only on partitional clustering and model-based clustering for mixed datasets. Balaji and Lavanya [13] present a short review paper on mixed data clustering. Many important mixed data clustering algorithms and research themes are omitted from in the paper. The paper also does not discuss the challenges or the future directions in this area. The review paper by Miyamoto *et al.* [14] discusses only the basic concepts of clustering, no mixed data clustering algorithm is described in the paper. The published literature review on mixed data clustering show several drawbacks:

- Most of these papers fail to identify concrete research themes or taxonomy to pave the way for performing systematic research in the field.
- None of these papers are comprehensive in their literature survey; thus, their scope is limited.
- Some papers focus on specific types of algorithms, whereas others review general algorithms without providing detailed insights and challenges.
- Most of these papers do not identify major application areas where mixed data clustering is relevant.
- The majority of these papers ignore important practical issues such as data availability, scalability of algorithms, big data challenges, and interpretability.
- Many papers do not focus on the future development of the field and does not provide guidelines to make progress.

The literature review presented in this paper attempts to avoid the drawbacks listed above and aims to contribute to the enhancement of knowledge in the field.

III. TAXONOMY FOR MIXED DATA CLUSTERING

In recent years, there has been a surge in the popularity of mixed data clustering algorithms because many real-world datasets contain both numeric and categorical features. Mixed data clustering can be performed in several ways, depending on the process involved in clustering the data points. However, there exists no unified framework to structure the research being undertaken in this field.

In this section, we present a new taxonomy to facilitate the study of state-of-the-art mixed data clustering algorithms. This taxonomy identifies five major research themes of clustering algorithms – *partitional*, *hierarchical*, *model-based*, *neural network-based*, and *other*. The 'other' category encompasses several minor groups of clustering algorithms that either do no fit into the other major research themes or have not been extensively studied. Therefore, we combine these emerging methods into a single broad research theme. A few clustering algorithms may belong to more than one research themes identified by the taxonomy; however, we take great care to place them in the most appro-

TABLE 2. Taxonomy for the study of mixed data clustering algorithms.

#	Research Themes	Research Papers
1	Partitional	Huang [9], [10], Ahmad and Dey [6], Huang *et al.* [15], Modha and Spangler [16], Chen and He [17], Ren *et al.* [18], Ji *et al.* [19], Sangam and OM [20], Roy and Sharma [21], Wang *et al.* [22], Wei *et al.* [23], Zhao *et al.* [24], Chiodi *et al.* [25], Kaeem *et al.* [26], Jang *et al.* [27], Barcelo-Rico and Jose-Luis [28], Wang *et al.* [22], Wei *et al.* [23], Cheng and Leu [29], Ahmad and Dey [30], Ji *et al.* [31], Kuri-Morales*et al.* [32], Ji *et al.* [33], Chen *et al.* [34], Wangchamhan *et al.* [35], Lakhsmi et al [36], Ahmad and Hashmi [37], Liang *et al.* [38], Yao *et al.* [39]
2	Hierarchical	Philips and Ottaway [40], Li and Biswas [41], Chiu *et al.* [42], Hsu *et al.* [43], , Hsu and Chen [44], Hsu and Huang [45], Shih *et al.* [46], Lim *et al.* [47], Chae *et al.* [48]
3	Model-based	Cheeseman and Stutz [49], Everitt [50], Moustaki and Papageorgiou [51], Browne and McNicholas [52], Andreopoulos *et al.* [53],Hunt and Jorgensen [54], Lawrence and Krzanowski [55], McParland and Gormley [56], Saâdaoui *et al.* [57], McParland [58], Rajan and Bhattacharya [59], Tekumalla *et al.* [60], Marbac *et al.* [61],Foss *et al.* [62], Doring *et al.* [63], Chatzis [64], Pathak and Pal [65]
4	Neural network-based	Devaraj and Punithavalli [66], Hsu [67], Hsu and Lin [68], [69],Tai and Hsu [70], Chen *et al.* [71],del Coso *et al.* [72], Noorbehbahani *et al.* [73], Lam *et al.* [74],Hsu and Huang [45]
5	Other	Luo *et al.* [75], David and Averbuchb [76], Niu *et al.* [77], Ahmad and Dey [78], Jia and Cheung [79], Plant and Böhm [80], Du *et al.* [81], [82], Liu et a. [83], Milenova and Campos [84], Mckusick and Thompson [85], Reich and Fenves [86], Ciaccio *et al.* [87], Sowjanya and Shashi [88], Frey and Dueck [89], Zhang and Gu [90], He *et al.* [91], He *et al.* [92], Hai and Susumu [93], Zhao *et al.* [94], Böhm *et al.* [95], Behzadi *et al.* [96], Plant [97], Li and Ye [98], Cheung and Jia [99], Sangam and Om [100], Lin *et al.* [101], Sangam and Om [102], Yu *et al.* [103]

priate thematic area of research. Table 2 shows the proposed taxonomy with five different type of research themes for clustering mixed data, along with the relevant research works that is reviewed in the subsequent sections.

A. PARTITIONAL CLUSTERING

The most studied research theme in mixed data clustering comes from the family of partitional clustering algorithms. Most of these algorithms share characteristics with partitional algorithms developed for pure numeric data (for example K-means [104]), or pure categorical data (for example K-modes [105]) or their variants. The general idea of these algorithms is to define

(i) a cluster center that can represent categorical features and numeric features
(ii) a distance measure that can combine numeric and categorical features, and
(iii) a cost function, which is minimized iteratively, that can handle mixed data.

Combining the above three ideas, most of the partitional clustering algorithms optimize the following cost function iteratively,

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

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

An important reason for the early adoption and widespread adaptability of partitional algorithms is that they are linear in the number of data points, scales well to large datasets and can be adapted to parallelization frameworks (for example MapReduce). Below, we review several key partitional algorithms to cluster mixed data.

Huang [9], [10] proposes the K-prototypes clustering algorithm for mixed datasets using 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 [5] is not a good representative of the similarity between feature values for a pair of multi-valued categorical feature values. The reason is that Hamming distance gives the distance between two categorical values as only 0 or 1 depending upon whether two features values are same or different. Hence, this measure cannot correctly capture the distance between two differing feature values. For example, in Table 1, the Hamming distance between feature values *Teaching* and *Medical* may not be the same as the distance between feature values *Teaching* and *Sportsman*. However, the Hamming distance will suggest otherwise and give a value of 0 in both cases.

Ahmad and Dey [6] propose a new cost function and a distance measure to address these problems. They calculate the similarity between two feature values of a categorical feature from the data. The similarity depends upon the co-occurrence of these feature values with feature values of other features. Weights of numeric features are also calculated in this method such that more significant features are given greater weights. A novel frequency-based representation of cluster center is proposed for categorical features, whereas the mean is used to for numeric features. It is shown that their proposed clustering algorithm performs better than the K-prototypes clustering algorithm.

Huang *et al.* [15] extend the K-prototypes clustering algorithm to propose the W-K-prototypes clustering algorithm. In each iteration, the feature weights are updated 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 the K-prototypes algorithm [9], [10]. Zhao *et al.* [24] use the frequency of feature values for categorical features to define the cluster centers. The Hamming distance measure was used to compute the distance for categorical features, whereas mean values are used for numeric features. They show improved clustering results in comparison to the K-prototypes algorithm [9], [10].

Modha and Spangler [16] 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 to have two feature spaces; one consisting of numeric features and the other with categorical features. Each numeric feature is linearly scaled (by subtracting by the mean and dividing by the standard deviation) and 1-in-q representation for each q-ary categorical feature is used. The squared Euclidean distance is used for numeric features, whereas the cosine distance is used for categorical features. No comparative study with other clustering algorithms is presented in the paper.

Chen and He [17] use the distance measure suggested by Ahmad and Dey [6] to propose a data clustering algorithm for data streams with mixed numeric 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 to cluster the data. The method uses two parameters: decay factor and dense threshold. Decay factor defines the significance of historical data to the current cluster whereas the dense threshold is used to distinguish between dense and sparse micro-clusters. The parameter optimization is a potential problem with the method.

Ren *et al.* [18] use the cluster centers proposed by Ahmad and Dey [6] to develop another mixed data clustering algorithm. Euclidean distance for numeric features and Hamming distance for categorical features are used to compute the similarity between the cluster center and a data point, with a Gaussian kernel function applied to the total distance. Ji *et al.* [19] combine the definition of cluster center [6] with the significance of feature [15] to propose a new cost function. The significance of a feature is initially selected randomly, followed by an update to its value with each iteration. The random selection of the significance of a feature can worsen the problem of random initialization of the cluster center [1], [106] because it would lead to different results in different runs.

Sangam and Om [20] propose a new distance measure for the K-prototypes clustering algorithms. The weightage Hamming distance is proposed for categorical features, this is based on the frequency of feature values in different clusters. The Minkowski distance measure is used to compute the distance for numeric features. The proposed method outperforms the original K-prototypes clustering algorithm.

Roy and Sharma [21] extend the fast genetic K-means clustering technique (FGKA) [107] for mixed data. The algorithm minimizes the total within-cluster variation. They use the distance measure proposed by Ahmad and Dey [6] in their algorithm. They claim that the algorithm performs better than the FGKA algorithm [107]; however, they do not explain the modification made in FGKA (which can handle only numeric data) to allow mixed data.

Chiodi [25] propose an iterative partitional clustering algorithm for mixed data, which is motivated by the K-means clustering algorithm. They propose a cost function which computes the mean diversity of the data points in a cluster with respect to all of the features. The Euclidean distance measure is used for a numeric feature and the Hamming distance is used for categorical features. Mean values are used for numeric features and the frequency distribution is used for categorical values in clusters. The algorithm is applied to the andrological dataset. Kacem *et al.* [26] propose parallelization of the K-prototypes clustering method [9] to handle large mixed datasets, this algorithm uses the MapReduce framework [108] for parallelization. Jang *et al.* [27] use a grid-based indexing technique to develop grid-based K-prototypes algorithm that speeds up K-prototypes algorithm. The experiments carried out using a spatial dataset consisting of numeric and categorical features show that the proposed method takes less time than the original K-prototypes algorithm. Table 3 summarizes different K-means-type algorithms for mixed data clustering.

The other partitional approach to mixed data clustering is to first convert a mixed dataset into a numeric dataset and then apply traditional K-means clustering to it. Barcelo-Rico and Jose-Luis [28] develop a method that uses polar or spherical coordinates to codify categorical features into numeric features and then uses K-means clustering on the new numeric datasets. Their method outperforms K-modes clustering algorithms and K-prototypes clustering method. Wang *et al.* [22] propose the context-based coupled representation for mixed datasets. The interdependence of numeric features and the interdependence of categorical features are computed separately and then, the interdependence across the numeric and categorical features is computed. These relationships form the numeric representation for mixed-type data points. The 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.* [23] propose a mutual information-based trans-

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

Algorithm	Center Definition	Distance Measure
Huang [9], [10]	Mean values for numeric features, mode values for categorical data	Euclidean distance for numeric features, Hamming distance for categorical features
Ahmad and Dey [6]	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.* [15]	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 [9], [10]
Zhao *et al.* [24]	Mean values for numeric features, proportional frequency-based center for categorical features	Euclidean distance for numeric features, Hamming distance for categorical features
Modha and Spangler [16]	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.* [19]	Center as proposed by Ahmad and Dey [6]	Weights are calculated by the method suggested by Huang *et al.* [15], squared Euclidean distance is used for numeric features, Hamming distance is used for categorical features
Ran *et al.* [18]	Center as proposed by Ahmad and Dey [6]	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 [28]	Coding is based on polar or spherical coordinates
Wang *et al.* [22]	Context-based coupled relationship for mixed data
Wei *et al.* [23]	Mutual information (MI)-based unsupervised feature transformation

formation method for unsupervised features that can convert categorical features into numeric features, which are then clustered by using K-means clustering algorithm. Table 4 summarizes the clustering methods that first convert the mixed data to numeric data and then apply the K-means clustering on the new numeric data.

Constraint-based clustering [109] groups similar data points into several clusters under certain user constraints: for example. that two given data points should belong to the same cluster. Cheng and Leu [29] propose a constrained K-prototypes clustering algorithm that simultaneously handles user constraints and mixed data. The algorithm extends the K-prototypes clustering algorithm [9] by adding a constrained function to the cost function of the K-prototypes.

Fuzzy clustering represent those approaches in which a data point can belong to more than one cluster with different degrees (or probabilities) of membership [110]. Various fuzzy clustering algorithms have been proposed for mixed data based on partitional clustering. Ahmad and Dey [30] use a dynamic probabilistic distance measure to determine the weights of numeric features and distances between 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 [111] to develop a fuzzy C-means (FCM) clustering algorithm [112], [113] for mixed data. Ji *et al.* [31] propose a fuzzy clustering method for mixed data by combining the similarity measure proposed by Ahmad and Dey [6] with the cluster center definition suggested by El-Sonbaty and Ismail [111]. Kuri-Morales *et al.* [32] propose a strategy for the assignment of a numeric value to a categorical value. First, a mixed dataset is converted into a pure numeric dataset and then fuzzy C-means clustering algorithm is used.

Partitional clustering algorithms for numeric and categorical data (for example K-means and K-modes) suffer from several drawbacks, such as cluster center initialization [1], [106] and the prior knowledge of the number of clusters [104]. Because of their conceptual similarity, these issues also exist in their counterparts for mixed datasets. In the next subsections, we review relevant literature that covers these issues.

1) CLUSTER CENTER INITIALIZATION

Cluster center initialization is a well-known problem with partitional clustering algorithms [1], [106], [114]. In these algorithms, initial cluster centers are usually selected randomly this may lead to different clustering outcomes on different runs of the algorithm. Therefore, data mining researchers may find it difficult to rely on such clustering outcomes.

Ji *et al.* [33] propose an algorithm to create initial cluster centers for K-means-type algorithms for mixed datasets. They introduce the idea of the centrality of data points, which uses the concept of neighbor-set. The centrality and distances are used to compute the initial cluster centers. However, their algorithm has quadratic complexity, in contrast to the linear time complexity of K-means-type clustering algorithms.

Using density peaks [115], Chen *et al.* [34] propose an 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-type algorithms. Wangchamhan *et al.* [35] combine a search algorithm, League Championship Algorithm [116], with the K-means clustering algorithm to identify the initial cluster centers. They apply Gower's distance measure [117] to find the distance between a data point and a cluster center. Parameter selection is a problem with this approach. Lakhsmi *et al.* [36] use the crow optimization method to compute the initial cluster centers for the K-prototypes clustering algorithm. This algorithm outperforms the K-prototypes clustering algorithm with random initial cluster centers. The selection of parameters in crow optimization is an important step; the same clustering results may not be produced by using different parameters.

Hashmi [37] combines the distance measure and the definition of centers for mixed data proposed by Ahmad and Dey [6] with the cost function of K-harmonic clustering [118] to extend K-harmonic clustering to mixed data. Their results indicate that their method is robust to the selection of initial cluster centers as compared to other K-means clustering type algorithms for mixed datasets. Zheng *et al.* [119] combine an evolutionary algorithm (EA) with the K-prototypes clustering algorithm [9]. The global searching ability of EA makes the proposed algorithm less sensitive to cluster initialization.

2) NUMBER OF CLUSTERS

Most of the partitional clustering algorithms for numeric and categorical data work under the assumption that the number of clusters is known in advance. This number may be either computed by other algorithms, derived from the domain, or user-defined. However, many of these methods may not guarantee that the chosen number of clusters corresponds to the natural number of clusters in the data. The same problem exists for partitional algorithms for mixed data.

Liang *et al.* [38] propose a cluster validity index to discover the number of clusters for mixed data clustering. This 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 [120]. For numeric features, a corresponding category utility function proposed by Mirkin [121] is used. Each component is given a weight depending upon the number of categorical and numeric features and the total number of features. The cluster validity index is computed for different number of clusters. The number of clusters that maximizes the cluster validity index is chosen as the optimal number of clusters. 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 [122] for numeric features and complement entropy [123] for categorical features are used to determine the worst cluster. The method is used with the K-prototypes method [9]. The algorithm is successful in finding the number of clusters in various datasets. These datasets have predefined classes and the number of the classes is taken as the number of clusters in the datasets. Yao *et al.* [39] extend the algorithm [38] by adding a method to find the initial clusters to avoid the cluster initialization problem. However, the method to find initial clusters is based on density estimation which makes the method quadratic. The comparative study suggests that the original method [38] may produce different number of clusters in different runs whereas the proposed method produces the same number of clusters. The experiment shows that the method is successful in predicting the correct number of clusters in datasets.

Rahman and Islam [124] combine genetic algorithm optimization [125] and the 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 [126] 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 methods create a hierarchy of clusters organized in a top to down (or bottom to up) order. To create clusters, the hierarchical algorithms need both of the following:

(i) Similarity matrix - This is constructed by finding the similarity between each pair of mixed data points. The choice of similarity metric (to construct a similarity matrix) influences the shape of the clusters,
(ii) Linkage criterion – This determines the distance between sets of observations as a function of the pairwise distances between observations.

Most hierarchical clustering algorithms have a large time complexity of $O(n^3)$ and requires $O(n^2)$ memory, where n is the number of data points. Below, we review several hierarchical clustering algorithms that have been developed to handle mixed data.

Philip and Ottaway [40] use Gower's similarity measure [117] 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 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 similarity between the same feature values is 1, whereas if the difference between the values is the maximum possible difference (the difference between maximum and minimum values of the feature) the similarity is 0. The sum of the similarity values for all numeric features is the similarity for two data points in a numeric feature space. The similarity in the categorical feature space and the numeric feature space are added to compute the similarity between two data points. Hierarchical agglomerative clustering is then used to create clusters. Fang *et al.* [42] develop a similarity measure to compute the similarity between two clusters for mixed data. This similarity measure is related to the decrease in the log-likelihood function when two clusters are merged. Zhang *et al.* [127] combine the BIRCH clustering algorithm, which uses hierarchical clustering algorithm, with their proposed similarity measure to develop a clustering algorithm that can handle mixed datasets. Li and Biswas [41] propose similarity-based agglomerative clustering (SBAC) algorithm for mixed data clustering. SBAC uses the Goodall similarity measure [128] and applies a hierarchical agglomerative approach to build cluster hierarchies.

Hsu *et al.* [43] propose a distance measure based on a concept hierarchy consisting of concept nodes and links [129], [130]. 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 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 algorithm [2] is applied to a distance matrix to obtain the clusters. Domain knowledge is required to make distance hierarchies for categorical features, and is non-trivial in many cases. Hsu and Chen [44] propose a new similarity measure to cluster mixed data. The algorithm uses variance for computing the similarity of numeric values. For similarity between categorical values, they [44] utilizes entropy with distance hierarchies [43]. The similarities are then aggregated to compute the similarity matrix for a mixed dataset. Incremental clustering is used on the similarity matrix to obtain the clusters. In an extended work, Hsu and Huang [45] apply an adaptive resonance theory network (ART) to cluster data points by using the distance hierarchies as the input of the network. A better interpretation of clusters is possible with the proposed algorithm as compared to the K-prototypes algorithm. Shih *et al.* [46] convert categorical features of a mixed dataset into numeric features by using frequencies of co-occurrence of categorical feature values. The dataset with all numeric features is then clustered by using a hierarchical agglomerative clustering algorithm [2].

Lim *et al.* [47] partition the data into two parts: categorical data and numeric data. The two types of data are clustered separately. The clustering results are combined by using a weighted scheme to obtain a similarity matrix. The agglomerative hierarchical clustering method is applied on the similarity matrix to obtain the final clusters. Gower's similarity measure assigns equal weights to both types of features in computing the similarity between two data points. The similarity matrices may be dominated by one feature type. Chae and Yang [48] assign weights to the different feature types to overcome this problem. Improved clustering results are shown with these weighted similarity matrices.

Table 5 summarizes the different hierarchical clustering methods for mixed data that were discussed in this section.

C. MODEL-BASED CLUSTERING

Model-based clustering methods assume that a data point matches a model, which in many cases, is a statistical distribution [132]. The models are generally user-defined, so they are prone to yielding undesirable clustering outcomes if inappropriate models (or their parameters) are chosen. Model-based clustering algorithms are generally slower than partitional algorithms [132]. Next, we review several model-based clustering algorithms for mixed data.

AUTOCLASS [49] performs clustering by integrating finite mixture distribution and Bayesian methods with prior distribution of each feature. AUTOCLASS can cluster data containing both categorical and numeric features. Everitt [50] 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. Because of high computational cost, the method is only useful for datasets containing very few categorical features. To overcome this problem, Lawrence and Krzanowski [55] extend the homogeneous Conditional

TABLE 5. Hierarchical clustering algorithms for mixed datasets.

Algorithm	Similarity measure for a similarity matrix	Clustering algorithm
Philip and Ottaway [40]	Gower's similarity Matrix [117]	Agglomerative hierarchical clustering method
Chiu *et al.* [42]	Probabilistic model by using a log-likelihood function	BIRCH algorithm [127]
Li and Biswas [41]	Goodall similarity measure [128]	Agglomerative hierarchical clustering with group- average method
Hsu *et al.* [43]	Distance hierarchy by using concept hierarchy [129], [130]	Agglomerative hierarchical clustering
Hsu and Chen [44]	Variance for numeric features and entropy with distance hierarchies [43] for categorical features	Incremental clustering
Hsu and Huang [45]	Similarity measure proposed by Hsu and Chen [44]	Adaptive resonance theory network [131]
Shih *et al.* [46]	Convert categorical features into numeric features	Hierarchical agglomerative clustering algorithm [2]
Lim *et al.* [47]	Two similarity matrices: one for categorical data and one for numeric data	Agglomerative hierarchical clustering method
Chae *et al.* [48]	Modified Gower's similarity matrix	Agglomerative hierarchical clustering method

Gaussian model to the finite mixture case, to compute maximum likelihood estimates for the parameters in a sample population. They suggest that their method works for an arbitrary number of features.

Moustaki and Papageorgiou [51] use a latent class mixture model for mixed data clustering. Categorical features are converted to binary features by a 1-in-q representation. A multinomial distribution is used for categorical features and a normal distribution is used for a numeric features. Features are considered independent in each cluister. The algorithm has been applied to an archaeological dataset. Browne and McNicholas [52] propose a mixture of latent features model for clustering, the expectation-maximization (EM) framework [133] is used for model fitting. Andreopoulos *et al.* [53] present a clustering algorithm, Bi-level clustering of mixed categorical and numeric data types (BILCOM) for mixed datasets. The algorithm uses categorical data clustering to guide the clustering of numeric data. Hunt and Jorgensen [54], [134], [135] 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. A 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 [134].

The ClustMD method [56] 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 the parameters for ClustMD. A Monte Carlo EM algorithm [136] is used for datasets having categorical features. This method can model both numeric and categorical features; however, it becomes computationally expensive as the number of features increases. To overcome this problem, McParland *et al.* [137] 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 the Gibbs sampling algorithm. To select the optimal model, they also propose an approximate Bayesian Information Criterion-Markov chain Monte Carlo criterion. They show that the method works well on a mixed medical dataset consisting of high-dimensional numeric phenotypic features and categorical genotypic features. Saâdaoui *et al.* [57] propose a projection of the categorical features on the subspaces spanned by numeric features; an optimal Gaussian mixture model is obtained from the resulting principal component analysis regressed subspaces.

Rajan and Bhattacharya [59] present a clustering algorithm based on Gaussian mixture copulas[1] that can model dependencies between features and can be applied to datasets having numeric and categorical features. Their method outperforms other clustering algorithms on a variety of datasets. Tekumalla *et al.* [60] use the concept of vines copulas[2] for mixed data clustering, they propose an inferencing algorithm to fit those vines on the mixed data. A dependency-seeking multi-view clustering that uses a Dirichlet process mixture of vines is developed [60]. Marbac *et al.* [61] 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 select the model by using two information criteria: the Bayesian information criterion [139] and integrated completed likelihood criterion [140]. The Bayesian inference is performed by using a Metropolis-within-Gibbs sampler. Foss *et al.* [62] develop a semi-parametric method, KAy-means for MIxed LArge data (KAMILA), 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 [135]. Like the K-means clustering algorithm, no strong parametric assumptions are made for numeric features in the KAMILA algorithm. KAMILA uses the properties of Gaussian-multinomial mixture models to balance the effects of numeric and categorical features without specifying weights.

Doring *et al.* [63] 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 finding the cluster prototypes. The inverse of the probability that a data point occurs in a cluster is used to define the distance between the cluster center and the data point. Chatzis [64] proposes 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 [141]. Pathak and Pal [65] combine fuzzy, probabilistic and collaborative clustering in a framework for mixed data clustering. Fuzzy clustering is used to cluster numeric data portion of the mixed data, whereas mixture models [3], [64] are used to cluster categorical data portion. Collaborative clustering [142] is used to find the common cluster sub-structures in the categorical and numeric data.

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

D. NEURAL NETWORK-BASED CLUSTERING

Most of the research on clustering mixed data using neural networks is focused on using self organizing maps (SOM) [143] and adaptive resonance theory (ART) [74] approaches. A SOM [143], [144] is a neural network that is used to non-linearly project a dataset onto a lower-dimensional feature space so that cluster analysis can be performed in the new fea-

[1] 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'' [138].

[2] Vine copulas provide a flexible way of pair-wise dependency modeling using hierarchical collections of bivariate copulas, each of which can belong to any copula family thereby capturing a wide variety of dependencies [60].

TABLE 6. Model-based clustering algorithms for mixed datasets.

Algorithm	Model
Cheeseman and Stutz [49]	Bayesian methods
Everitt [50]	Model-based clustering with the use of thresholds for the categorical features.
Lawrence and Krzanowski [55]	Extension of homogeneous conditional Gaussian model to the finite mixture situation.
Moustaki and Papageorgiou [51]	Latent class mixture model.
Browne and McNicholas [52]	A mixture of latent variables model with the expectation-maximization framework. [133].
Andreopoulos *et al.* [53]	Pseudo-Bayesian process with categorical data clustering to guide the clustering of numeric data.
Hunt and Jorgensen [54], [134], [135]	A finite mixture of multivariate distributions is fitted to data.
McParland and Gormley [56]	A latent variable model.
McParland *et al.* [58]	Bayesian finite mixture model.
Saadaoui *et al.* [57]	A projection of the categorical features on the subspaces spanned by numeric features and then the application of Gaussian Mixture Model.
Rajan and Bhattacharya [59]	Gaussian mixture copula.
Tekumalla1 *et al.* [60]	Vine copulas and Dirichlet process mixture of vines.
Marbac [61]	A mixture model of Gaussian copulas.
Foss *et al.* [62]	K-means algorithm and Gaussian-multinomial mixture models

ture space. ART is based on the theory of how the brain learns to categorize autonomously and predict in a dynamic world [145]. The key aspect of ART's predictive power is its ability to carry out fast, incremental, and stable unsupervised and supervised learning in response to a changing world [145]. Both the traditional SOM-based and ART-based clustering methods can handle numeric features, however they cannot be used directly for categorical features. Categorical features are first transformed into binary features, which are then treated as numeric features [66], [74].

Hsu [67] develops a generalized SMO model to compute the 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 (numeric and categorical) feature space. Visualization-induced SMO [146] preserves the structure of data in the new low-dimensional space better than SMO. Hsu and Lin [68] 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 [69] 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 the weakness that it has predefined fixed-size map; to improve its flexibility, growing SMO is proposed [147]. Growing SMO starts with a small map that grows with training data. Tai and Hsu [70] integrate generalized SMO with growing SMO to develop a clustering algorithm for mixed datasets. Chen and Marques [71] propose a clustering algorithm based on SMO, using the Hamming distance for categorical features and the Euclidean distance for numeric features. This method has the problem that it gives more weight to categorical features, to overcome this problem del Coso *et al.* [72] 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 [71]. Noorbehbahani *et al.* [73] propose an incremental mixed-data clustering algorithm which uses a self-organizing incremental neural network algorithm [148]. They also propose a new distance measure in which the distance between two categorical values depend on the frequencies of those features. The co-occurrence of feature values [6], which may affect the accuracy of the distance measure, is not considered.

Lam *et al.* [74] use an unsupervised feature learning approach to obtain a sparse representation of mixed datasets. A fuzzy adaptive resonance theory (ART) approach [149] is used to create new features. First, fuzzy ART approach is used to create prototypes of the dataset, which are then employed as mixed features encoder to map individual data points to the new feature space. They use K-means clustering algorithm to cluster data points in the new feature space. Hsu and Huang [45] use ART to create a similarity matrix that can be used to cluster data points by using hierarchical clustering.

E. OTHER

In the previous sections, we summarized major contributions on the four prominent research themes adopted by researchers for clustering mixed data. However, several new sub-themes and research directions have emerged in recent years. As many of these new research directions have not been explored enough, we combine them under one umbrella theme named '*Other*'. Many of these new types of clustering algorithms may not fit within the realms of the more established research themes as discussed in previous sections.

Spectral clustering techniques [150] perform dimensionality reduction by using eigenvalues of the similarity matrix of the data. Thereafter, the clustering is performed in fewer dimensions. First a similarity matrix is computed, and then a spectral clustering algorithm [150] is applied to this similarity matrix to obtain clusters. Luo *et al.* [75] propose a similarity measure by using a clustering ensemble technique. In this measure, the similarity of two data points is computed separately for numeric and categorical features. The two similarities are added to obtain the similarity between two data points. Spectral clustering is used on the similarity matrix to obtain the clusters. David and Averbuch [76] propose a clustering algorithm, SpectralCAT, which uses categorical spectral clustering to cluster mixed datasets. The algorithm automatically transforms the numeric features to categorical values. This is performed by finding the optimal transformation according to the Calinski and Harabasz [151] index. A spectral clustering method is then applied to the transformed data [76]. Niu *et al.* [77] present a clustering algorithm for mixed data, in which the similarity matrices for numeric and categorical features are computed separately. Coupling relationships of features are used to compute similarity matrices. Both matrices are combined by weighted summation to compute the similarity matrix for the mixed data. This algorithm is applied to find the clusters for a web-based learning system data, The results suggest that the method outperforms the K-prototypes clustering algorithm and the SpectralCAT algorithm [76].

Subspace clustering [152] seeks to discover clusters in different subspaces within a dataset. Ahmad and Dey [78] use a distance measure [6] for the mixed data with a cost function for subspace clustering [153] to develop a K-means-type clustering algorithm, which can produce subspace clustering of mixed data. Jia and Cheung [79] 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 the most appropriate number of clusters automatically. Plant and Böhm [80] develop a clustering technique, interpretable clustering of numeric and categorical objects (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 [154]. 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. The algorithm demands that all values of categorical features involved in a dependency with some numeric features must have a unique numeric data distribution.

Density-based clustering methods assume that clusters are defined by dense regions in the data space, separated by less dense regions [155]. Du *et al.* [81] and Ding *et al.* [82] propose a new distance measure for mixed data clustering, in which they assign a weight to each categorical feature. They combine this distance measure with a density peaks clustering algorithm [115] to cluster mixed datasets. However, the selection of different parameters makes it difficult to use in practice. Liu *et al.* [83] propose a density-based clustering algorithm for mixed datasets. Ester *et al.* [155] extend the DBSCAN algorithm to mixed datasets. Entropy is used to compute the distance measure for mixed datasets. Milenova [84] use orthogonal projections to cluster mixed datasets. These orthogonal projections are used to find high-density regions in the input data space. Du *et al.* [156] propose a density-based clustering method for mixed datasets. Datasets can be divided into three categories depending upon the ratio of the number of categorical features and the number of numeric features. Different mathematical models are suggested for these categories. First, numeric features are used to create clusters categorical features are used to create clusters, and finally, these clusters are combined to obtain the final clusters.

Conceptual clustering [157] generates a concept description for each generated cluster. Generally, conceptual clustering methods generate hierarchical category structures. COBWEB [157] uses a category utility (CU) measure [120] to define the relation between groups or clusters. As the CU measure can only handle categorical features, the CU measure is extended to handle numeric features for mixed data clustering. COBWEB3 [85] integrates the original COBWEB algorithm with the method presented in CLASSIT [158] to deal with numeric features in the CU measure. With this method, it is assumed that numeric feature values are normally distributed. To overcome the problem of normal distribution assumption, a new method ECOBWEB [86], which uses the probability distribution of the average value for a feature, is presented.

Ciaccio [87] extends the well-separated partition definition [159] to propose a non-hierarchical clustering algorithm for mixed data, which can analyze large amount of data in the presence of missing values. Sowjanya and Shashi [88] 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, which are updated as new data points join the clusters. A cluster center is defined, for a categorical feature, by using the mode of the categorical values of data points present in the cluster. For a numeric feature, the mean of the values of the data points present in a cluster is used to represent the center of the cluster. However, it is not clear in the paper which distance measure is used to cluster data points.

Frey and Dueck [89] propose an affinity propagation clustering (APC) algorithm that uses message passing. Zhang and Gu [90] extend this method by combining the distance measure proposed by Ahmad and Dey [6] with the APC algorithm. Accurate clustering results are achieved with this method. He *et al.* [91] extend the Squeezer algorithm [160] which works for pure categorical datasets for clustering mixed data. In one of the versions, the numeric features are discretized to convert them to categorical features and then Squeezer algorithm is applied to the new categorical data. In another work, He *et al.* [92] divide the mixed data into two parts: pure numeric features and pure categorical features. A graph partitioning algorithm is used to cluster numeric data, whereas categorical data is clustered by using the Squeezer algorithm. The clustering results are combined and treated as categorical data, which is clustered by using the Squeezer algorithm to get the final clustering results. Hai and

Susumu [93] parallelize the clustering algorithm proposed by He *et al.* [91] to handle large datasets.

Zhao *et al.* [94] present an ensemble method, which creates base clustering models in sequence, for mixed dataset. The clustering models are created so that they have large diversity. The first base clustering model is created by a random partition of data points. In each run, a clustering model is generated and each data point is checked to find whether changing its cluster membership will decrease the value of a proposed optimization function. The complexity of this algorithm is quadratic. As the start of the proposed algorithm is random, the final clustering results may be different with different initial random clusters.

Böhm *et al.* [95] propose a parameter-free clustering algorithm, INTEGRATE, for mixed data. The algorithm is based on the concept of MDL [154]. This allows the balancing of the effects of numeric and categorical features. INTEGRATE is scalable to large datasets. Behzadi *et al.* [96] propose a distance hierarchy to compute the distances for mixed datasets. A modified DBSCAN clustering algorithm is used to cluster the data and the MDL principle is used for clustering without specifying parameters.

Plant [97] proposes a clustering algorithm, scale-free dependency clustering (Scenic), 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 [161]. The clusters are then obtained in the new low-dimensional embedding.

Li and Ye [98] 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 distance measures. Mohanavalli and Jaisakthiusing [210] use chi-square statistics for computing the weight of each feature of mixed data. The Euclidean distance for numeric features and the Hamming distance for categorical features along with these weights are used to compute the distances. The authors did not describe about the clustering algorithm used in their paper.

Cheung and Jia [99] present a general clustering framework that uses the concept of similarity between data point and cluster, and propose a unified similarity metric for mixed datasets. Accordingly, they propose an iterative clustering algorithm that finds the number of clusters automatically. Sangam and Om [100] present a sampling-based clustering algorithm for mixed datasets. The algorithm has two steps: first, a sample of data points is used for clustering, and then 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. In their method, the clustering algorithm presented by Cheung and Jia [99] is used in the first step.

Lin *et al.* [101] present a tree-ensembles clustering algorithm, CRAFTER, for clustering high-dimensional mixed datasets. First, a random subset of data points is drawn and the random forests clustering algorithm [162] is applied. The clustered data points are used to train tree classifiers. These trained tree-ensembles are used to cluster all of the data points.

Sangam and Om [102] present a clustering algorithm for time-evolving data streams. They propose a window-based method to detect concept drift. The data characteristics of features in the current sliding window are compared with those of the previous sliding window; the frequency is used for a categorical feature and the mean and standard deviation are used for a numeric feature. A similarity difference that exceeds the user-defined threshold indicates a concept drift.. The clustering algorithm proposed by Cheung and Jia [99] is used to show the results.

Three-way clustering deals with three decisions; a data point certainly belongs to a cluster, a data point may belong to a cluster (uncertain) and a data point certainly does not belong to a cluster. Yu *et al.* [103] propose a three-way clustering algorithm for mixed datasets. They propose a new distance measure to compute the distance between two data points. A tree-based distance measure is proposed for categorical features. The difference between normalized feature values is used for numeric features. The algorithm uses a mixed data clustering algorithm and thresholds The references are missing from the paper, so it cannot be studied in detail. Xiong and Yu [163] extend this work and present an adaptive three-way clustering algorithm for mixed datasets which can produce three-way clustering without thresholds.

IV. ANALYSIS OF THE SURVEY

The previous section reviews the majority of the key clustering algorithms around five broad research themes for mixed data. Some of the newer and less developed areas of research are combined into the '*Other*' theme. We also observed that few studies encompass more than one research theme (for example combining ideas from partitional and neural network-based clustering). However, we noted that algorithms based on partitional clustering are mostly favored by researchers and practitioners, because these algorithms are:

- simpler in interpretation and implementation;
- linear in the number of data objects; so they scale well with big data application;
- easily adaptable to parallel architectures, making them more practical to apply to big data problems.

Despite these advantages, finding an appropriate similarity measure and cost function to handle mixed data remains a challenge in partitional clustering algorithms. Nonetheless, these algorithms work well in practice. The hierarchical, model-based, neural network-based, and other clustering approaches may provide better clustering outcomes; however, either they suffer from nonlinear time or space complexity or they involve making assumptions about the data distribution that may not hold in real-world scenarios. These reasons further impede progress in non-partitional clustering algorithms.

Research developments are taking place to address the problems of traditional clustering algorithms, such as the problems of cluster initialization and the number of desired clusters (for partitional algorithms) and the selection of the proper model and reasonable parameter assumptions (for model-based clustering). New trends in clustering, including subspace clustering, spectral clustering, clustering ensembles, big data clustering, and data stream clustering have been suggested for mixed datasets.

A major issue in evaluating these clustering algorithms is the choice of performance metric. In an ideal clustering scenario, class labels are not available-this is, in fact, the rationale behind performing unsupervised learning. In the absence of class labels, evaluating the performance of clustering algorithms is not straightforward. Typically, the datasets that are used to demonstrate mixed data clustering results have class labels, which 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 a clustering algorithm. Therefore, as ground truth labels are present (but are not used to perform clustering), many performance measures have been used in the literature, including F-measure, normalized mutual information, and rand index [2]. However, in our survey, we found that clustering accuracy has been the most commonly used criterion for evaluating 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 the i^{th} cluster and their 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 literature survey, we found a lack of comparison between competitive clustering algorithms. Part of the problem is the choice of different datasets by various algorithms. It emerged that some of the popular datasets used by many researchers to evaluate their algorithms are: Heart (Cleveland), Heart (Statlog), and Australian Credit data. However, these datasets are relatively small in size, and may not be representative of real-world datasets and complex problems.

In the next section, we present several publicly available software packages for performing mixed data clustering and list some major application areas.

V. SOFTWARE AND APPLICATIONS

A. SOFTWARE

As the field of mixed data clustering progresses, many researchers have made software packages and libraries available for use by the wider community. The majority of these software packages are available in R [164]. The K-prototypes clustering algorithm [9] is available in R [165]. The ClustMD package in R [166] is the implementation of model-based clustering for mixed data [58]. Gower's similarity matrix [117] is implemented in R. The similarity matrix can be used with the partitioning around medoids tools in R or the hierarchical clustering tools to obtain final clusters [167]. ClustOfVar [168] is an R package for clustering that can handle mixed datasets. Both a 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 [169]. An implementation of the KAMILA [62] clustering algorithm is available in R [170]. The mixed data clustering algorithm by Macbar *et al.* [61] is implemented in R [171]. Ahmad and Dey [6] mixed data clustering algorithm is available in Matlab [172]. A K-means-type clustering algorithm that can deal with mixed datasets is implemented in Matlab, using feature discretization [173]. MixtComp is a C++ implementation of model-based clustering of mixed data [174].

B. MAJOR APPLICATION AREAS

Most of the real world applications contain mixed data. Some of these application areas are (but not limited to) health, marketing, business, finance, social studies. Below, we present a list of major application areas where mixed-data clustering is mostly applied.

1) HEALTH AND BIOLOGY

McParland and Gormley [56], and McParland *et al.* [58] 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.* [175] use mixed data clustering to study people who died of cancer between 1994 and 2006 in Hijuelas. Storlie *et al.* [176] develop model-based clustering for mixed datasets with missing feature values to cluster autism spectrum disorder. Researchers have used various types of clustering approaches for mixed data for heart disease [6], [16], [41], [78], occupational Medicine [57], [177], digital mammograms [178], acute inflammations [31], [65], [97], age of abalone snails [97], human life span [179], dermatology [80], medical diagnosis [98], toxicogenomics [180], genetic Regulation, analysis of biomedical datasets, [53] and cancer Samples Grouping [181].

2) BUSINESS AND MARKETING

Hennig and Liao [7] apply mixed data clustering techniques for socio-economic stratification by using 2007 US survey data of consumer finances. Kassi *et al.* [182] develop a 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 [6], [15],

[16], [41], [78], income prediction (adult data) [16], [19], [45], marketing research [183], customer behavior discovery [184], customer segmentation and catalog marketing [44], customer behavior pattern discovery [185], motor insurance [186] and construction management [29].

3) OTHER APPLICATIONS

Moustaki and Papageorgiou [51] apply mixed data clustering in archaeometry for classifying archaeological findings into groups. Philip and Ottaway [40] use mixed data clustering to cluster Cypriot hooked-tang weapons. Chiodi [25] use mixed data clustering for andrological data [25]. Iam-On and Boongoen use mixed data clustering for student dropout prediction in a Thai university. Mixed data clustering has also been used in teaching assistant evaluation [38], [74], class examination [135], petroleum recovery [74], intrusion detection [18], [98], [188], forest cover type [26], online learning systems [77], automobiles [80], printing process delays [28] and country flags mining [189].

VI. IMPACT AREAS, CHALLENGES AND OPEN RESEARCH QUESTIONS

A. IMPACT AREAS

As discussed in Section V-B, mixed data clustering algorithms have been applied in various application domains. We believe that employing mixed data clustering in multiple domains is very important; however, we argue that the areas of health and business informatics will have more impact because they attempt to solve real-world problems that are related to people.

1) HEALTH INFORMATICS

The majority of the data for health applications are based on either electronic health records (EHR) [190] or sensors [191]. EHR data can contain a patient's medical history, diagnoses, medications, treatment plans, immunization dates, allergies, radiology images, and laboratory and test results [192]. EHR is a great resource to allow the deployment of evidence-based supervised and unsupervised machine learning tools to make decisions about patients' care. Therefore, EHR data is a good example of mixed data with high-impact real-world applications. Data from sensors can be either numeric (for example, motion or physiology) or categorical (for example, door open or closed). These datasets are important in building machine learning driven applications for rehabilitation, assessment of medical conditions, and detection and prediction of health-related events [193], [194]. Application of mixed data clustering on these datasets is important in identifying medical conditions among people with disability, morbidity, and cognitive disorders. Clustering on these diverse datasets can also help in performing sex and gender based research, and vulnerable populations and older adults.

2) BUSINESS ANALYTICS

Business analytic is another domain in which a large number of mixed datasets are created. Market research is an important area in this domain. Analysis of customer datasets that contain categorical features (for example type of a customer, preference, and income group) and numeric features (for example, age and the number of transactions) provide managers with insights about the customer behavior [183]. Credit card data analysis is used to predict the financial health of an individual. Generally, credit card datasets are mixed datasets. Various clustering algorithms have been applied on mixed credit datasets [6], [16]. The financial statements of a company are analyzed to assess the company's financial health; the datasets consisting of categorical features (for example, the type of the company, products and the region of the company) along with numeric features (for example, financial ratios) present better information about a company. People analytics [195] is an emerging area: companies are interested in knowing about present and future employees to improve their productivity and satisfaction. Employee datasets consisting of categorical features (education, department, and job type) and numeric features (age, years in job, and salary) can capture information about employees better than datasets containing only one type of feature.

B. CHALLENGES

In the previous sections, we mentioned several technical challenges for mixed-data clustering algorithms. We now summarize those challenges for each research theme of the taxonomy with detailed ideas for future research directions.

1) PARTITIONAL CLUSTERING

As noted previously, one of the reasons of widespread usage of partitional clustering algorithm for mixed data is their linear time complexity with respect to the number of data points. However, the notion of center may not be clearly defined for these algorithms. Therefore, combining numeric and categorical centers to initialize these algorithms is not straightforward and it requires more research to obtain a good representation of the concept of cluster center. Another related aspect of these algorithms is finding the similarity between data objects and cluster centers. The literature suggests the development of several distance measures [6], [9], [15]; however, the scale by which numeric and categorical distances are combined is not clear. Among the available similarity measures, there is no unanimous winner and this specific area needs more research.

The literature review suggests that cluster center initialization may help in learning consistent and robust clusters. Several methods have also been proposed for that purpose [33], [34], [37]; however, there is no method that is both computationally inexpensive and gives consistent results in different runs. Finding good initial clusters is the key to the success of these algorithms and must be treated as an active area of research. Similarly, estimating the number of clusters in a mixed dataset is an important and challenging problem. Identifying a number of clusters that is close to the natural number of clusters in the dataset can enhance our understanding of not only the dataset but also the underlying problem.

2) HIERARCHICAL CLUSTERING

The majority of hierarchical clustering algorithms rely on calculating a similarity matrix, from which clusters can be constructed. However, the similarity matrix depends on a good definition of similarity or distance. As stated above, the distance between two mixed data objects is not self-explanatory and requires more research.

3) MODEL-BASED CLUSTERING

As observed in the literature review, the majority of the model-based mixed data clustering algorithms suffer from high model complexity. The selection of an appropriate model is an important step in model-based clustering. There are two types of features in mixed datasets, so the selection of models for these two types of features is a challenging task. Modeling the conditional dependency between categorical and numeric features is another challenge. Selecting appropriate parametric assumptions is a difficult problem for model-based clustering. As mixed datasets have categorical features, which are not continuous, this problem is more serious for model-based clustering. As there are two types of features, identifying important features for distinguishing clusters presents a difficult task. These drawbacks may turn out be an obstacle to employing such powerful methods on large datasets to solve real-world problems. Therefore, significant effort is needed to develop models that can work with fewer parameters and offer lower model complexity.

4) NEURAL NETWORK-BASED CLUSTERING

The majority of the research work on clustering mixed data using neural networks is centered around SOM and ART. The SOM methods may lead to poor topological mappings and may not be able to match the structure of the distribution of the data [196]. The ART models are typically governed by differential equations and have high computational complexity [196]. There are several other areas of traditional neural network-based clustering that can be adapted for mixed datasets: for example, adaptive subspace SOM, ARTMAP, and learning vector quantization [196].

C. OPEN RESEARCH QUESTIONS AND GUIDELINES

In this section, we highlight several open questions that may be relevant to the different types of clustering algorithms discussed in the proposed taxonomy.

- Cluster ensembles have shown great promise for clustering numeric datasets by significantly improving the results of the base clustering algorithm [197], [198]. However, more research is desirable for developing robust cluster ensemble methods for mixed datasets.
- It is well known that real-world mixed data is imperfect; missing values among features is one such major issue that may impair the capabilities of many existing clustering algorithms. One plausible approach is to first impute missing mixed data values [199] and perform existing clustering methods. The other approach is to develop clustering algorithms that can handle missing data in their objective function [134]. However, the development of, and comparison between, these two types of competitive approaches has not been investigated much and this may require attention from the research community to solve real-world problems.
- Various mixed datasets in application areas such as medical and socio-economics contain uncertain data because of improper data acquisition methods or inherent problems in the data acquisition. In our review, we could not find methods that can handle these types of datasets. Clustering uncertain mixed datasets is an important research direction, with applications in many domains.
- Few researchers have developed methods for converting a mixed dataset to a pure numeric dataset, so that clustering algorithms meant for pure numeric datasets can be employed [23], [74]. This is indeed a new perspective on the difficult problem of mixed data clustering. We further note that transformation of mixed data to numeric data does not come without loss of information. Therefore, it is an open question to the research community to develop algorithms that can reduce the adverse effects of data transformation.
- Clustering with deep learning approaches is an emerging area of research [200], [201]. The objective (loss) function of deep learning clustering methods is primarily composed of the deep network loss and the clustering loss. Therefore, these methods differ according to the network architecture (for example, autoencoder, variational autoencoder, or generative adversarial network) or the type of clustering method (such as partitional or hierarchical). However, these methods are mostly aimed at numeric datasets; there is a great opportunity to explore mixed data clustering alongside deep learning methods.
- As datasets increase in size and domains become complex, the majority of successful machine learning algorithms lose their interpretability and may be treated as a black box. Mixed data clustering algorithms are no exception. The idea of clustering models that are easy to explain is attractive to practitioners, such as clinicians, business analysts, geologists, and biologists. Interpretable models can assist them in making informed decisions. Unfortunately, only a few researchers have explored this area of developing interpretable mixed data clustering methods to address critical aspects of the models: for example, why a certain set of data points forms a cluster or how different clusters can be distinguished from each other [202]. Novel research in this area will produce outcomes outside the realms of the

research community. Many clustering algorithms may benefit from reducing the dimensions of multivariate mixed data, as a result of reducing their execution time and model complexity. There has been recent research in the field of feature selection for mixed data [203], [204]; however, combining such results with clustering has not been much explored. Selecting a subset of relevant features has the potential to enhance the interpretability of clustering algorithms as well.

- Another repercussion of big data is ensuring the scalability of clustering algorithms, to make them useful in real-world scenarios. Parallelization of mixed data clustering algorithms is a viable approach [26] to allow them to scale with increasing data size and maintain linear time complexity (especially for partitional clustering). Active research in this area is required to keep the field in synchronization with big data challenges. Similarly, developing fast and accurate online clustering algorithms to handle large streams of mixed data requires attention to address shortcomings. These include low clustering quality, evaluation of new concepts and concept drift in the underlying data, difficulties in determining cluster centers, and poor ability to deal with outliers [17].
- Subspace clustering is a viable approach to cluster large quantities of high-dimensional mixed data, though the large data problem in itself is very challenging. The extension of other subspace clustering approaches, for example, grid-based methods for mixed datasets [205] is key to development of clustering algorithms for high-dimensional mixed datasets. In subspace clustering, a data point can belong to more than one cluster and subspaces are axis-parallel [206]. Research on adapting other subspace clustering approaches that have been developed for numeric datasets, such as correlation clustering [207], should be extended to mixed data clustering. In particular, using the correlation between numeric and categorical features to create subspaces is an innovative research area.
- Integration of domain knowledge into clustering is an important research area as this can improve the clustering accuracy and cluster interpretation. Constrained clustering is an approach to handling problems of this type. Constrained clustering for iterative partitional clustering methods has been proposed for mixed datasets [109]; however, there has been no research work on the application of constrained clustering to other approaches of clustering, such as hierarchical and density-based clustering. With the availability of large domain knowledge, there is a need to develop clustering algorithms for mixed data that can utilize this knowledge to create more accurate and interpretable clusters.
- Several clustering algorithms require user-defined parameters. Therefore, the final clustering results are strongly dependent on these parameters, which include the number of desired clusters and initial clusters for iterative partitional clustering algorithms and the model selection for model-based clustering. Some efforts have been made to develop parameter-free clustering algorithms for mixed datasets [95], [96]; however, research in this field is quite open-ended.
- Spectral clustering produces good results and does not require strong assumptions about the statistics of the clusters. Spectral clustering has been used to cluster mixed datasets [75], [77]. The similarity matrix is the first step of spectral clustering. Each spectral clustering method for mixed datasets develops its own similarity matrix [75], [77]. A large number of similarity measures are available for mixed datasets. A detailed study is required to understand which similarity measures are more useful for spectral clustering.
- As a pure unsupervised machine learning paradigm, true labels should not be present during clustering. Thus, evaluating the performance of clustering algorithms in this situation is not straight forward. However, in certain experimental scenarios, true labels may be present, and they can be used for matching with clustering labels. In the literature review, we also observed that *accuracy* has been reported by many researchers as a performance metric for clustering algorithms (see Section IV). A major problem with using *accuracy* or other confusion matrix-based performance measures is that they assume a direct correspondence between true and clustering labels. However, clustering labels are arbitrary and matching them with true labels is non-trivial. With small data size and a small number of natural clusters, this technique of matching true and clustering labels may be feasible, with support from domain knowledge. However, *accuracy* will be difficult to comprehend *accuracy* as the number of clusters and data points increase. Therefore, in experimental scenarios where the true labels are known, performance metrics such as adjusted rand index, normalized mutual information, homogeneity, completeness, and the V-measure [208] are more relevant and should be widely adopted. For real-world clustering problems where the true labels are not be present, performance indexes such as the silhouette coefficient, Calinski-Harabaz Index, and Davies-Bouldin Index [209] should be used.
- A ubiquitous problem that has been highlighted in this literature review is that the majority of clustering algorithms tested their methods on a few publicly available datasets. Moreover, several researchers showed results on datasets that were not available to the wider community. We believe that creating a community-based mixed data repository not only provides opportunities to compare existing clustering methods and set benchmarks but also encourages the development of new algorithms at a faster pace. Furthermore, we believe that sharing and contributing clustering algorithms' code in the public domain, by means such as R packages, Python libraries, and Java classes is useful for quickly comparing and testing existing and new methods. As discussed

in Section V, some software packages have been made public; more effort will certainly benefit the research community.

In this paper, we identified five major research themes for the study of mixed data clustering and presented a comprehensive state-of-the-art survey of literature within them. We discussed the challenges and future directions within each research theme, and discussed several high-impact application areas, open research questions, and guidelines for making progress in the field. This survey paper should guide researchers to develop an in-depth understanding of the field of mixed data clustering and help generate new ideas to make significant contributions to solve real-world problems.

REFERENCES

[1] S. S. Khan and A. Ahmad, "Cluster center initialization algorithm for K-means clustering," *Pattern Recognit. Lett.*, vol. 25, no. 11, pp. 1293–1302, 2004.
[2] A. Jain and R. Dubes, *Algorithms for Clustering Data*. Upper Saddle River, NJ, USA: Prentice-Hall, 1988.
[3] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer-Verlag, 2008.
[4] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. San Mateo, CA, USA: Morgan Kaufmann, 2005.
[5] S. Boriah, V. Chandola, and V. Kumar, "Similarity measures for categorical data: A comparative evaluation," in *Proc. SIAM Int. Conf. Data Mining*, 2008, pp. 243–254.
[6] A. Ahmad and L. Dey, "A k-mean clustering algorithm for mixed numeric and categorical data," *Data Knowl. Eng.*, vol. 63, no. 2, pp. 503–527, 2007.
[7] C. Hennig and T. F. Liao, "How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification," *J. Roy. Stat. Soc. C, Appl. Stat.*, vol. 62, no. 3, pp. 309–369, 2013.
[8] I. I. Morlini and S. Zani, "Comparing approaches for clustering mixed mode data: An application in marketing research," in *Data Analysis and Classification*. Berlin, Germany: Springer, 2010, pp. 49–57.
[9] Z. Huang, "Extensions to the K-means algorithm for clustering large data sets with categorical values," *Data Mining Knowl. Discovery*, vol. 2, no. 3, pp. 283–304, Sep. 1998.
[10] Z. Huang, "Clustering large data sets with mixed numeric and categorical values," in *Proc. 1st Pacific Asia Knowl. Discovery Data Mining Conf.* Singapore: World Scientific, 1997, pp. 21–34.
[11] M. van de Velden, A. I. D'enza, and A. Markos, "Distance-based clustering of mixed data," *WIREs Comput. Statist.*, 2018.
[12] A. H. Foss, M. Markatou, and B. Ray, "Distance metrics and clustering methods for mixed-type data," *Int. Stat. Rev.*, 2018.
[13] K. Balaji and K. Lavanya, "Clustering algorithms for mixed datasets: A review," *Int. J. Pure Appl. Math.*, vol. 18, no. 7, pp. 547–556, 2018.
[14] S. Miyamoto, V. Huynh, and S. Fujiwara, "Methods for clustering categorical and mixed data: An overview and new algorithms," in *Integrated Uncertainty in Knowledge Modelling and Decision Making*. Cham, Switzerland: Springer, 2018, pp. 75–86.
[15] J. Z. Huang, M. K. Ng, H. Rong, and Z. Li, "Automated variable weighting in k-means type clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 657–668, May 2005.
[16] D. S. Modha and W. S. Spangler, "Feature weighting in k-means clustering," *Mach. Learn.*, vol. 52, no. 3, pp. 217–237, Sep. 2003.
[17] J.-Y. Chen and H.-H. He, "A fast density-based data stream clustering algorithm with cluster centers self-determined for mixed data," *Inf. Sci.*, vol. 345, pp. 271–293, Jun. 2016.
[18] M. Ren, P. Liu, Z. Wang, and X. Pan, "An improved mixed-type data based kernel clustering algorithm," in *Proc. 12th Int. Conf. Natural Comput., Fuzzy Syst. Knowl. Discovery (ICNC-FSKD)*, 2016, pp. 1205–1209.
[19] J. Ji, T. Bai, C. Zhou, C. Ma, and Z. Wang, "An improved k-prototypes clustering algorithm for mixed numeric and categorical data," *Neurocomputing*, vol. 120, pp. 590–596, Nov. 2013.
[20] R. K. Sangam and H. Om, "An equi-biased k-prototypes algorithm for clustering mixed-type data," *Sādhanā*, vol. 43, no. 3, p. 37, 2018.
[21] D. K. Roy and L. K. Sharma, "Genetic k-means clustering algorithm for mixed numeric and categorical data sets," *Int. J. Artif. Intell.*, vol. 1, no. 2, pp. 23–28, 2010.
[22] C. Wang, C. Chi, W. Zhou, and R. Wong, "Coupled interdependent attribute analysis on mixed data," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 1861–1867.
[23] 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*, vol. 17, no. 3, pp. 1535–1548, 2015.
[24] W. Zhao, W. Dai, and C. Tang, "K-centers algorithm for clustering mixed type data," in *Proc. 11th Pacific-Asia Conf. Adv. Knowl. Discovery Data Mining (PAKDD)*, 2007, pp. 1140–1147.
[25] M. Chiodi, "A partition type method for clustering mixed data," *Riv. Stat. Appl.*, vol. 2, no. 2, pp. 135–147, 1990.
[26] M. A. B. Kacem, C.-E. B. N'cir, and N. Essoussi, "Mapreduce-based k-prototypes clustering method for big data," in *Proc. IEEE Int. Conf. Data Sci. Adv. Anal. (DSAA)*, Oct. 2015, pp. 1–7.
[27] H.-J. Jang, B. Kim, J. Kim, and S. Jung, "An efficient grid-based k-prototypes algorithm for sustainable decision-making on spatial objects," *Sustainability*, vol. 10, no. 8, p. 2614, 2018.
[28] F. Barcelo-Rico and J.-L. Diez, "Geometrical codification for clustering mixed categorical and numerical databases," *J. Intell. Inf. Syst.*, vol. 39, no. 1, pp. 167–185, 2012.
[29] Y. M. Cheng and S. S. Leu, "Constraint-based clustering and its applications in construction management," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 5761–5767, 2009.
[30] A. Ahmad and L. Dey, "Algorithm for fuzzy clustering of mixed data with numeric and categorical attributes," in *Distributed Computing and Internet Technology*. Berlin, Germany: Springer, 2005, pp. 561–572.
[31] J. Ji, W. Pang, C. Zhou, X. Han, and Z. Wang, "A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data," *Knowl.-Based Syst.*, vol. 30, pp. 129–135, Jun. 2012.
[32] A. Kuri-Morales, D. Trejo-Baños, and L. E. Cortes-Berrueco, "Clustering of heterogeneously typed data with soft computing—A case study," in *Proc. Mexican 10th Int. Conf. Artif. Intell.*, vol. 2. New York, NY, USA: Springer-Verlag, 2011, pp. 235–248.
[33] 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," *Appl. Math. Inf. Sci.*, vol. 9, no. 6, pp. 2933–2942, 2015.
[34] J. Chen, X. L. Xiang, H. Zheng, and X. Bao, "A novel cluster center fast determination clustering algorithm," *Appl. Soft Comput.*, vol. 57, pp. 539–555, Aug. 2017.
[35] 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 Syst. Appl.*, vol. 90, pp. 146–167, Dec. 2017.
[36] K. Lakshmi, N. V. Karthikeyani, S. Shanthi, and S. Parvathavarthini, "Clustering mixed datasets using k-prototype algorithm based on crow-search optimization," in *Developments and Trends in Intelligent Technologies and Smart Systems*. Hershey, PA, USA: IGI Global, 2017, pp. 129–150.
[37] A. Ahmad and S. Hashmi, "K-harmonic means type clustering algorithm for mixed datasets," *Appl. Soft Comput.*, vol. 48, pp. 39–49, Nov. 2016.
[38] J. Liang, X. Zhao, D. Li, F. Cao, and C. Dang, "Determining the number of clusters using information entropy for mixed data," *Pattern Recognit.*, vol. 45, no. 6, pp. 2251–2265, Jun. 2012.
[39] X. Yao, S. Ge, H. Kong, and H. Ning, "An improved clustering algorithm and its application in wechat sports users analysis," *Procedia Comput. Sci.*, vol. 129, pp. 166–174, Dec. 2018.
[40] G. Philip and B. S. Ottaway, "Mixed data cluster analysis: An illustration using Cypriot hooked-tang weapons," *Archaeometry*, vol. 25, no. 2, pp. 119–133, 1983.
[41] C. Li and G. Biswas, "Unsupervised learning with mixed numeric and nominal data," *IEEE Trans. Knowl. Data Eng.*, vol. 14, no. 4, pp. 673–690, Jul./Aug. 2002.
[42] 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 *Proc. 7th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2001, pp. 263–268.
[43] C.-C. Hsu, C.-L. Chen, and Y.-W. Su, "Hierarchical clustering of mixed data based on distance hierarchy," *Inf. Sci.*, vol. 177, no. 20, pp. 4474–4492, 2007.

[44] C.-C. Hsu and Y.-C. Chen, "Mining of mixed data with application to catalog marketing," *Expert Syst. Appl.*, vol. 32, no. 1, pp. 12–23, 2007.

[45] C. C. Hsu and Y. P. Huang, "Incremental clustering of mixed data based on distance hierarchy," *Expert Syst. Appl.*, vol. 35, no. 3, pp. 1177–1185, 2008.

[46] M. Shih, J. Jheng, and L. Lai, "A two-step method for clustering mixed categroical and numeric data," *Tamkang J. Sci. Eng.*, vol. 13, no. 1, pp. 11–19, 2010.

[47] J. Lim, J. Jun, S. Kim, and D. McLeod, "A framework for clustering mixed attribute type datasets," in *Proc. 4th Int. Conf. Emerg. Databases (EDB)*, 2012, pp. 1–12.

[48] S.-S. Chae, J.-M. Kim, and W.-Y. Yang, "Cluster analysis with balancing weight on mixed-type data," *Commun. Stat. Appl. Methods*, vol. 13, no. 3, pp. 719–732, 2006.

[49] P. Cheeseman and J. Stutz, "Bayesian classification (AutoClass): Theory and results," in *Advances in Knowledge Discovery and Data Mining*. Cambridge, MA, USA: MIT Press, 1996, pp. 153–180.

[50] B. S. Everitt, "A finite mixture model for the clustering of mixed-mode data," *Statist. Probab. Lett.*, vol. 6, no. 5, pp. 305–309, 1988.

[51] I. Moustaki and I. Papageorgiou, "Latent class models for mixed variables with applications in archaeometry," *Comput. Statist. Data Anal.*, vol. 48, no. 3, pp. 659–675, 2005.

[52] R. P. Browne and P. D. McNicholas, "Model-based clustering, classification, and discriminant analysis of data with mixed type," *J. Stat. Planning Inference*, vol. 142, no. 11, pp. 2976–2984, 2012.

[53] B. Andreopoulos, A. An, and X. Wang, "Bi-level clustering of mixed categorical and numerical biomedical data," *Int. J. Data Mining Bioinf.*, vol. 1, no. 1, pp. 19–56, 2006.

[54] L. Hunt and M. Jorgensen, "Mixture model clustering of data sets with categorical and continuous variables," in *Information, Statistics and Induction in Science*. Singapore: World Scientific, 1996, pp. 375–384.

[55] C. J. Lawrence and W. J. Krzanowski, "Mixture separation for mixed-mode data," *Statist. Comput.*, vol. 6, no. 1, pp. 85–92, 1996.

[56] D. McParland and I. C. Gormley, "Model based clustering for mixed data: ClustMD," *Adv. Data Anal. Classification*, vol. 10, no. 2, pp. 155–169, 2016.

[57] 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 Trans. Nanobiosci.*, vol. 14, no. 7, pp. 707–715, Oct. 2015.

[58] D. McParland, C. M. Phillips, 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," *Stat. Med.*, vol. 36, no. 28, pp. 4548–4569, 2017.

[59] V. Rajan and S. Bhattacharya, "Dependency clustering of mixed data with Gaussian mixture copulas," in *Proc. 25th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2016, pp. 1967–1973.

[60] L. S. Tekumalla, V. V. Rajan, and C. Bhattacharyya, "Vine copulas for mixed data: Multi-view clustering for mixed data beyond meta-Gaussian dependencies," *Mach. Learn.*, vol. 106, no. 9, pp. 1331–1357, 2017.

[61] M. Marbac, C. Biernacki, and V. Vandewalle, "Model-based clustering of gaussian copulas for mixed data," *Commun. Statist.-Theory Methods*, vol. 46, no. 23, pp. 11635–11656, 2017.

[62] A. Foss, M. Markatou, B. Ray, and A. Heching, "A semiparametric method for clustering mixed data," *Mach. Learn.*, vol. 105, no. 3, pp. 419–458, 2016.

[63] C. Doring, C. Borgelt, and R. Kruse, "Fuzzy clustering of quantitative and qualitative data," in *Proc. IEEE Annu. Meeting Fuzzy Inf. Process. (NAFIPS)*, vol. 1, Jun. 2004, pp. 84–89.

[64] 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 Syst. Appl.*, vol. 38, no. 7, pp. 8684–8689, 2011.

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

[66] H. P. Devaraj and M. Punithavalli, "An integrated framework for mixed data clustering using self organizing map," *J. Comput. Sci.*, vol. 7, no. 11, pp. 1639–1645, 2011.

[67] C.-C. Hsu, "Generalizing self-organizing map for categorical data," *IEEE Trans. Neural Netw.*, vol. 17, no. 2, pp. 294–304, Mar. 2006.

[68] C. Hsu and S. Lin, "Visualized analysis of multivariate mixed-type data via an extended self-organizing map," in *Proc. 6th Int. Conf. Inf. Technol. Appl. (ICITA)*, 2006, pp. 218–223.

[69] C.-C. Hsu and S.-H. Lin, "Visualized analysis of mixed numeric and categorical data via extended self-organizing map," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 1, pp. 72–86, Jan. 2012.

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

[71] N. Chen and N. C. Marques, "An extension of self-organizing maps to categorical data," in *Progress in Artificial Intelligence*. Berlin, Germany: Springer, 2005, pp. 304–313.

[72] C. del Coso, D. Fustes, C. Dafonte, F. J. Nóvoa, J. M. Rodríguez-Pedreira, and B. Arcay, "Mixing numerical and categorical data in a self-organizing map by means of frequency neurons," *Appl. Soft Comput.*, vol. 36, pp. 246–254, Nov. 2015.

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

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

[75] H. Luo, F. Kong, and Y. Li, "Clustering mixed data based on evidence accumulation," in *Advanced Data Mining and Applications*. Berlin, Germany: Springer, 2006, pp. 348–355.

[76] G. David and A. Averbuch, "SpectralCAT: Categorical spectral clustering of numerical and nominal data," *Pattern Recognit.*, vol. 45, no. 1, pp. 416–433, 2012.

[77] 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," *Math. Problems Eng.*, vol. 2015, 2015, Art. no. 747628.

[78] A. Ahmad and L. Dey, "A k-means type clustering algorithm for subspace clustering of mixed numeric and categorical datasets," *Pattern Recognit. Lett.*, vol. 32, no. 7, pp. 1062–1069, 2011.

[79] H. Jia and Y.-M. Cheung, "Subspace clustering of categorical and numerical data with an unknown number of clusters," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 8, pp. 3308–3325, Aug. 2018.

[80] C. Plant and C. Böhm, "Inconco: Interpretable clustering of numerical and categorical objects," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2011, pp. 1127–1135.

[81] M. Du, S. Ding, and Y. Xue, "A novel density peaks clustering algorithm for mixed data," *Pattern Recognit. Lett.*, vol. 97, pp. 46–53, Oct. 2017.

[82] 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," *Knowl.-Based Syst.*, vol. 133, pp. 294–313, Oct. 2017.

[83] X. Liu, Q. Yang, and L. He, "A novel DBSCAN with entropy and probability for mixed data," *Cluster Comput.*, vol. 20, no. 2, pp. 1313–1323, Jun. 2017.

[84] B. L. Milenova and M. M. Campos, "Clustering large databases with numeric and nominal values using orthogonal projections," Oracle Data Mining Technol., Oracle Corp., Tech. Rep., 2003. [Online]. Available: https://docs.oracle.com/database/121/DMCON/GUID-4CA9F5BD-8CA4-41A0-88D7-8C85F2CB816F.htm#DMCON059

[85] K. Mckusick and K. Thompson, "Cobweb/3: A portable implementation," NASA Ames Research Center, Mountain View, CA, USA, Tech. Rep. FIA-90-6-18-2, 1990.

[86] Y. Reich and J. S. Fenves, "The formation and use of abstract concepts in design," in *Concept Formation: Knowledge and Experience in Unsupervised Learning*, D. H. Fisher, Jr., M. J. Pazzani, and P. Langley, Eds. San Francisco, CA, USA: Morgan Kaufmann, 1991, pp. 323–353.

[87] A. Di Ciaccio, "MIXISO: A non-hierarchical clustering method for mixed-mode data," in *Advances in Classification and Data Analysis*. Berlin, Germany: Springer, 2001, pp. 27–34.

[88] A. M. Sowjanya and M. Shashi, "A cluster feature-based incremental clustering approach to mixed data," *J. Comput. Sci.*, vol. 7, no. 12, pp. 1875–1880, 2011.

[89] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, pp. 972–976, Feb. 2007.

[90] K. Zhang and X. Gu, "An affinity propagation clustering algorithm for mixed numeric and categorical datasets," *Math. Problems Eng.*, vol. 2014, 2014, Art. no. 486075.

[91] Z. He, X. Xu, and S. Deng, "Scalable algorithms for clustering large datasets with mixed type attributes," *Int. J. Intell. Syst.*, vol. 20, no. 10, pp. 1077–1089, 2005.

[92] Z. He, X. Xu, and S. Deng. (2005). "Clustering mixed numeric and categorical data: A cluster ensemble approach." [Online]. Available: https://arxiv.org/abs/cs/0509011
[93] N. T. M. Hai and H. Susumu, "Performances of parallel clustering algorithm for categorical and mixed data," in *Parallel and Distributed Computing: Applications and Technologies*. Berlin, Germany: Springer, 2005, pp. 252–256.
[94] X. Zhao, F. Cao, and J. Liang, "A sequential ensemble clusterings generation algorithm for mixed data," *Appl. Math. Comput.*, vol. 335, pp. 264–277, Oct. 2018.
[95] C. Böhm, S. Goebl, A. Oswald, C. Plant, M. Plavinski, and B. Wackersreuther, "Integrative parameter-free clustering of data with mixed type attributes," in *Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining*. Berlin, Germany: Springer, 2010, pp. 38–47.
[96] S. Behzadi, M. A. Ibrahim, and C. Plant, "Parameter free mixed-type density-based clustering," in *Database and Expert Systems Applications*. Cham, Switzerland: Springer, 2018, pp. 19–34.
[97] C. Plant, "Dependency clustering across measurement scales," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 361–369.
[98] X. Li and N. Ye, "A supervised clustering and classification algorithm for mining data with mixed variables," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 36, no. 2, pp. 396–406, Mar. 2006.
[99] Y.-M. Cheung and H. Jia, "Categorical-and-numerical-attribute data clustering based on a unified similarity metric without knowing cluster number," *Pattern Recognit.*, vol. 46, no. 8, pp. 2228–2238, 2013.
[100] R. S. Sangam and H. Om, "Hybrid data labeling algorithm for clustering large mixed type data," *J. Intell. Inf. Syst.*, vol. 45, no. 2, pp. 273–293, Oct. 2015.
[101] S. Lin, B. Azarnoush, and G. C. Runger, "CRAFTER: A tree-ensemble clustering algorithm for static datasets with mixed attributes and high dimensionality," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 9, pp. 1686–1696, Sep. 2018.
[102] R. S. Sangam and H. Om, "Equi-clustream: A framework for clustering time evolving mixed data," *Adv. Data Anal. Classification*, vol. 12, no. 4, pp. 973–995, Dec. 2018.
[103] H. Yu, Z. Chang, and B. Zhou, "A novel three-way clustering algorithm for mixed-type data," in *Proc. IEEE Int. Conf. Big Knowl. (ICBK)*, Aug. 2017, pp. 119–126.
[104] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probab.*, vol. 1, L. M. L. Cam and J. Neyman, Eds. Berkeley, CA, USA: Univ. California Press, 1967, pp. 281–297.
[105] S. S. Khan and S. Kant, "Computation of initial modes for k-modes clustering algorithm using evidence accumulation," in *Proc. 20th Int. Joint Conf. Artif. Intell.* San Mateo, CA, USA: Morgan Kaufmann, 2007, pp. 2784–2789.
[106] S. S. Khan and A. Ahmad, "Cluster center initialization algorithm for K-modes clustering," *Expert Syst. Appl.*, vol. 40, no. 18, pp. 7444–7456, 2013.
[107] Y. Lu, S. Lu, F. Fotouhi, Y. Deng, and J. S. Brown, "FGKA: A fast genetic k-means clustering algorithm," in *Proc. ACM Symp. Appl. Comput.*, New York, NY, USA, 2004, pp. 622–623.
[108] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," *Commun. ACM*, vol. 51, no. 1, pp. 107–113, 2008.
[109] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl, "Constrained k-means clustering with background knowledge," in *Proc. 18th Conf. Mach. Learn.*, 2001, pp. 577–584.
[110] M. S. Yang, "A survey of fuzzy clustering," *Math. Comput. Model.*, vol. 18, no. 11, pp. 1–16, 1993.
[111] Y. El-Sonbaty and M. A. Ismail, "Fuzzy clustering for symbolic data," *IEEE Trans. Fuzzy Syst.*, vol. 6, no. 2, pp. 195–204, May 1998.
[112] J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," *J. Cybern.*, vol. 3, no. 3, pp. 32–57, 1973.
[113] J. C. Bezdek, "Pattern recognition with fuzzy objective function," in *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York, NY, USA: Springer, 1981.
[114] S. S. Khan and A. Ahmad, "Computing initial points using density based multiscale data condensation for clustering categorical data," in *Proc. 2nd Int. Conf. Appl. Artif. Intell. (ICAAIP)*, 2003.
[115] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, Jun. 2014.
[116] A. H. Kashan, "League championship algorithm: A new algorithm for numerical function optimization," in *Proc. Int. Conf. Soft Comput. Pattern Recognit.*, Dec. 2009, pp. 43–48.
[117] J. C. Gower, "A general coefficient of similarity and some of its properties," *Biometrics*, vol. 27, no. 4, pp. 857–871, 1971.
[118] B. Zhang, "Generalized K-harmonic means–dynamic weighting of data in unsupervised learning," in *Proc. SIAM Int. Conf. Data Mining*. Philadelphia, PA, USA: SIAM, 2001, pp. 1–13.
[119] Z. Zheng, M. Gong, J. Ma, L. Jiao, and Q. Wu, "Unsupervised evolutionary clustering algorithm for mixed type data," in *Proc. IEEE Congr. Evol. Comput.*, Jul. 2010, pp. 1–8.
[120] M. Gluck, "Information, uncertainty and the utility of categories," in *Proc. 7th Annu. Conf. Cognit. Sci. Soc.* Irvine, CA, USA: Lawrence Erlbaum Associates, 1985, pp. 283–287.
[121] B. Mirkin, "Reinterpreting the category utility function," *Mach. Learn.*, vol. 45, no. 2, pp. 219–228, Nov. 2001.
[122] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Symp. Math. Statist. Probab.*, 1961, pp. 547–561.
[123] J. Liang, K. S. Chin, C. Dang, and R. C. M. Yam, "A new method for measuring uncertainty and fuzziness in rough set theory," *Int. J. Gen. Syst.*, vol. 31, no. 4, pp. 331–342, 2002.
[124] M. A. Rahman and M. Z. Islam, "A hybrid clustering technique combining a novel genetic algorithm with K-means," *Knowl.-Based Syst.*, vol. 71, pp. 345–365, Nov. 2014.
[125] M. Mitchell, *An Introduction to Genetic Algorithms* (Complex Adaptive Systems). Cambridge, MA, USA: MIT Press, 1998.
[126] M. A. Rahman and M. Z. Islam, "CRUDAW: A novel fuzzy technique for clustering records following user defined attribute weights," in *Proc. Data Mining Anal. (AusDM)*, Sydney, NSW, Australia, 2012, pp. 27–41.
[127] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: A new data clustering algorithm and its applications," *Data Mining Knowl. Discovery*, vol. 1, no. 2, pp. 141–182, 1997.
[128] D. W. Goodall, "A new similarity index based on probability," *Biometrics*, vol. 22, no. 4, pp. 882–907, 1966.
[129] J. Han and Y. Fu, "Dynamic generation and refinement of concept hierarchies for knowledge discovery in databases," in *Proc. AAAIWS 3rd Int. Conf. Knowl. Discovery Data Mining*, 1994, pp. 157–168.
[130] J. Han, Y. Cai, and N. Cercone, "Data-driven discovery of quantitative rules in relational databases," *IEEE Trans. Knowl. Data Eng.*, vol. 5, no. 1, pp. 29–40, Feb. 1993.
[131] G. A. Carpenter and S. Grossberg, "Adaptive resonance theory," in *Encyclopedia of Machine Learning and Data Mining*. Boston, MA, USA: Springer, 2010, pp. 22–35.
[132] V. Melnykov and R. Maitra, "Finite mixture models and model-based clustering," *Stat. Survey*, vol. 4, pp. 80–116, 2010.
[133] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc., B (Methodol.)*, vol. 39, no. 1, pp. 1–38, 1977.
[134] L. Hunt and M. Jorgensen, "Mixture model clustering for mixed data with missing information," *Comput. Statist. Data Anal.*, vol. 41, nos. 3–4, pp. 429–440, 2003.
[135] L. Hunt and M. Jorgensen, "Clustering mixed data," *Data Mining Knowl. Discovery*, vol. 1, no. 4, pp. 352–361, 2011.
[136] G. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*. Hoboken, NJ, USA: Wiley, 2008.
[137] D. McParland, C. M. Phillips, L. Brennan, H. M. Roche, and I. C. Gormley, "Clustering high-dimensional mixed data to uncover subphenotypes: Joint analysis of phenotypic and genotypic data," *Statist. Med.*, vol. 36, no. 28, pp. 4548–4569, 2017.
[138] R. B. Nelsen, *An Introduction to Copulas* (Lecture Notes in Statistics). Berlin, Germany: Springer-Verlag, 2006.
[139] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, 1978.
[140] C. Biernacki, G. Celeux, and G. Govaert, "Assessing a mixture model for clustering with the integrated completed likelihood," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 7, pp. 719–725, Jul. 2000.
[141] K. Honda and H. Ichihashi, "Regularized linear fuzzy clustering and probabilistic PCA mixture models," *IEEE Trans. Fuzzy Syst.*, vol. 13, no. 4, pp. 508–516, Aug. 2005.
[142] W. Pedrycz, "Collaborative fuzzy clustering," *Pattern Recognit. Lett.*, vol. 23, no. 14, pp. 1675–1686, Dec. 2002.
[143] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biol. Cybern.*, vol. 43, no. 1, pp. 59–69, 1982.

[144] T. Kohonen, M. R. Schroeder, and T. S. Huang, Eds., *Self-Organizing Maps*, 3rd ed. Berlin, Germany: Springer-Verlag, 2001.
[145] S. Grossberg, "Adaptive resonance theory: How a brain learns to consciously attend, learn, and recognize a changing world," *Neural Netw.*, vol. 37, pp. 1–47, Jan. 2013.
[146] H. Yin, "ViSOM—A novel method for multivariate data projection and structure visualization," *IEEE Trans. Neural Netw.*, vol. 13, no. 1, pp. 237–243, Jan. 2002.
[147] D. Alahakoon, S. K. Halgamuge, and B. Srinivasan, "Dynamic self-organizing maps with controlled growth for knowledge discovery," *IEEE Trans. Neural Netw.*, vol. 11, no. 3, pp. 601–614, May 2000.
[148] S. Furao and O. Hasegawa, "An incremental network for on-line unsupervised classification and topology learning," *Neural Netw.*, vol. 19, no. 1, pp. 90–106, Jan. 2006.
[149] G. A. Carpenter, S. Grossberg, and D. B. Rosen, "Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system," *Neural Netw.*, vol. 4, no. 6, pp. 759–771, 1991.
[150] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. 14th Int. Conf. Neural Inf. Process. Syst., Natural Synth. (NIPS)*, 2001, pp. 849–856.
[151] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Commun. Stat., Theory Methods*, vol. 3, no. 1, pp. 1–27, 1974.
[152] L. Parsons, E. Haque, and H. Liu, "Subspace clustering for high dimensional data: A review," *ACM SIGKDD Explorations Newslett.*, vol. 6, no. 1, pp. 90–105, Jun. 2004.
[153] 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. Knowl. Data Eng.*, vol. 19, no. 8, pp. 1026–1041, Aug. 2007.
[154] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.
[155] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowl. Discovery Data Mining (KDD)*, 1996, pp. 226–231.
[156] H. Du, W. Fang, H. Huang, and S. Zeng, "MMDBC: Density-based clustering algorithm for mixed attributes and multi-dimension data," in *Proc. IEEE Int. Conf. Big Data Smart Comput. (BigComp)*, Jan. 2018, pp. 549–552.
[157] D. H. Fisher, "Knowledge acquisition via incremental conceptual clustering," *Mach. Learn.*, vol. 2, no. 2, pp. 139–172, 1987.
[158] J. H. Gennari, P. Langley, and D. Fisher, "Models of incremental concept formation," *Artif. Intell.*, vol. 40, nos. 1–3, pp. 11–61, 1989.
[159] N. Jardine and R. Sibson, *Mathematical Taxonomy*. London, U.K.: Wiley, 1971.
[160] H. Zengyou, X. Xiaofei, and D. Shengchun, "Squeezer: An efficient algorithm for clustering categorical data," *J. Comput. Sci. Technol.*, vol. 17, no. 5, pp. 611–624, 2002.
[161] G. Michailidis and J. de Leeuw, "The Gifi system of descriptive multivariate analysis," *Statist. Sci.*, vol. 13, no. 4, pp. 307–336, Nov. 1998.
[162] T. Shi and S. Horvath, "Unsupervised learning with random forest predictors," *J. Comput. Graph. Statist.*, vol. 15, no. 1, pp. 118–138, Jun. 2006.
[163] J. Xiong and H. Yu, "An adaptive three-way clustering algorithm for mixed-type data," in *Foundations of Intelligent Systems*. Cham, Switzerland: Springer, 2018.
[164] RDC Team. (2008). R: A Language Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. [Online]. Available: http://www.R-project.org
[165] G. Szepannek. (2017). *clustMixType: K-Prototypes Clustering for Mixed Variable-Type Data*. Accessed: Jan. 28, 2018. [Online]. Available: https://cran.r-project.org/web/packages/clustMixType/index.html
[166] D. McParland and I. C. Gormley. (2017). *clustMD: Model Based Clustering for Mixed Data*. Accessed: Jan. 28, 2018. [Online]. Available: https://cran.r-project.org/web/packages/clustMD/index.html
[167] Wicked Good Data-R. (2016). *Clustering Mixed Data Types in R*. Accessed: Jan. 28, 2018. [Online]. Available: https://www.r-bloggers.com/clustering-mixed-data-types-in-r/
[168] M. Chavent, J. Saracco, V. Kuentz-Simonet, and B. Liquet, "ClustOfVar: An R package for the clustering of variables," *J. Stat. Softw.*, vol. 50, no. 13, pp. 1–16, 2012.
[169] M. Hummel, D. Edelmann, and A. Kopp-Schneider. (2017). *CluMix: Clustering and Visualization of Mixed-Type Data*. Accessed: Jan. 28, 2018. [Online]. Available: https://cran.r-project.org/web/packages/CluMix/index.html
[170] A. Foss and M. Markatou. (2016). *Kamila: Methods for Clustering Mixed-Type Data*. Accessed: Jan. 28, 2018. [Online]. Available: https://cran.r-project.org/web/packages/kamila/index.html
[171] M. Marbac, C. Biernacki, and V. Vandewalle. (2014). *Copules-Package: Mixed Data Clustering by a Mixture Model of Gaussian Copulas*. Accessed: Jan. 28, 2018. [Online]. Available: https://rdrr.io/rforge/MixCluster/man/Copules-package.html
[172] A. Alsahaf. (2016). *Mixed K-Means Package*. Accessed: Jan. 28, 2018. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/53489-amjams-mixed_kmeans
[173] C. Bock. (2016). *Mixed K-Means Clustering Algorithm With Variable Discretization*. Accessed: Jan. 28, 2018. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/55601-mixed-k-means-clustering-algorithm-with-variable-discretization
[174] C. Biernacki and V. Kubicki. (2016). *Mixtcomp Software for Full Mixed Data*. Accessed: Jan. 28, 2018. [Online]. Available: https://modal.lille.inria.fr/wikimodal/doku.php?id=mixtcomp
[175] 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 *Proc. 11th Conf. Artif. Intell. Med. (AIME)*, 2007, pp. 190–194.
[176] C. B. Storlie *et al.*, "Clustering and variable selection in the presence of mixed variable types and missing data," *Statist. Med.*, vol. 37, no. 19, pp. 2884–2899, 2018.
[177] 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 Trans. Nanobiosci.*, vol. 14, no. 7, pp. 707–715, Oct. 2015.
[178] S. M. Halawani, M. Alhaddad, and A. Ahmad, "A study of digital mammograms by using clustering algorithms," *J. Sci. Ind. Res.*, vol. 71, pp. 594–600, Sep. 2012.
[179] A. Kuri-Morales, L. E. Cortes-Berrueco, and D. Trejo-Baños, "Clustering of heterogeneously typed data with soft computing—A case study," in *Proc. 10th Int. Conf. Artif. Intell., Adv. Soft Comput. (MICAI)*. Berlin, Germany: Springer-Verlag, 2011, pp. 235–248.
[180] P. R. Bushel, "Clustering of mixed data types with application to toxicogenomics," Ph.D. dissertation, Dept. Bioinformatics, North Carolina State Univ., Raleigh, NC, USA, 2006.
[181] F. N. Z. Abidin and R. D. Westhead, "Flexible model-based clustering of mixed binary and continuous data: Application to genetic regulation and cancer," *Nucleic Acids Res.*, vol. 45, no. 7, p. e53, 2017.
[182] 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 *Proc. IEEE/ACS 12th Int. Conf. Comput. Syst. Appl. (AICCSA)*, Nov. 2015, pp. 1–6.
[183] I. Morlini and S. Zani, "Comparing approaches for clustering mixed mode data: An application in marketing research," in *Data Analysis and Classification*. Berlin, Germany: Springer, 2010, pp. 49–57.
[184] M. Cheng, Y. Xin, Y. Tian, C. Wang, and Y. Yang, "Customer behavior pattern discovering based on mixed data clustering," in *Proc. IEEE Int. Conf. Comput. Intell. Softw. Eng. (CiSE)*, Dec. 2009, pp. 1–4.
[185] M. Cheng, Y. Xin, Y. Tian, C. Wang, and Y. Yang, "Customer behavior pattern discovering based on mixed data clustering," in *Proc. Int. Conf. Comput. Intell. Softw. Eng.*, Dec. 2009, pp. 1–4.
[186] Z. Huang, "Clustering large data sets with mixed numeric and categorical values," in *Proc. 1st Pacific-Asia Conf. Knowl. Discovery Data Mining*, 1997, pp. 21–34.
[187] N. Iam-On and T. Boongoen, "Improved student dropout prediction in Thai University using ensemble of mixed-type data clusterings," *Int. J. Mach. Learn. Cybern.*, vol. 8, no. 2, pp. 497–510, 2017.
[188] N. Liu, "The research of intrusion detection based on mixed clustering algorithm," in *Computational Intelligence and Intelligent Systems*. Berlin, Germany: Springer, 2012, pp. 92–100.
[189] T. Li and Y. Chen, "A weight entropy k-means algorithm for clustering dataset with mixed numeric and categorical data," in *Proc. IEEE 5th Int. Conf. Fuzzy Syst. Knowl. Discovery (FSKD)*, vol. 1, Oct. 2008, pp. 36–41.
[190] P. B. Jensen, L. J. Jensen, and S. Brunak, "Mining electronic health records: Towards better research applications and clinical care," *Nature Rev. Genet.*, vol. 13, no. 6, pp. 395–405, 2012.
[191] S. S. Khan and J. Hoey, "Review of fall detection techniques: A data availability perspective," *Med. Eng. Phys.*, vol. 39, pp. 12–22, Jan. 2017.
[192] Office of the National Coordinator for Health Information Technology. (2018). *What is an Electronic Health Record (EHR)?* Accessed: Dec. 18, 2018. [Online]. Available: https://www.healthit.gov/faq/what-electronic-health-record-ehr

[193] S. S. Khan *et al.*, "Daad: A framework for detecting agitation and aggression in people living with dementia using a novel multi-modal sensor network," in *Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW)*, Nov. 2017, pp. 703–710.
[194] S. S. Khan, B. Ye, B. Taati, and A. Mihailidis, "Detecting agitation and aggression in people with dementia using sensors—A systematic review," *Alzheimer's Dementia*, vol. 14, no. 6, pp. 824–832, 2018.
[195] E. Houghton and M. Green, "People analytics: Driving business performance with people data," CIPD, London, U.K., Tech. Rep., 2018.
[196] K.-L. Du, "Clustering: A neural network approach," *Neural Netw.*, vol. 23, no. 1, pp. 89–107, 2010.
[197] J. Ghosh and A. Acharya, "Cluster ensembles," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 1, no. 4, pp. 305–315, 2011.
[198] S. Vega-Pons and J. Ruiz-Shulcloper, "A survey of clustering ensemble algorithms," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 25, no. 3, pp. 337–372, 2011.
[199] V. Audigier, F. Husson, and J. Josse, "A principal component method to impute missing values for mixed data," *Adv. Data Anal. Classification*, vol. 10, no. 1, pp. 5–26, 2016.
[200] E. Aljalbout, V. Golkov, Y. Siddiqui, and D. Cremers. (2018). "Clustering with deep learning: Taxonomy and new methods." [Online]. Available: https://arxiv.org/abs/1801.07648
[201] E. Min, X. Guo, Q. Liu, G. Zhang, J. Cui, and J. Long, "A survey of clustering with deep learning: From the perspective of network architecture," *IEEE Access*, vol. 6, pp. 39501–39514, 2018.
[202] C. Plant and C. Böhm, "INCONCO: Interpretable clustering of numerical and categorical objects," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2011, pp. 1127–1135.
[203] X. Zhang *et al.*, "Feature selection in mixed data: A method using a novel fuzzy rough set-based information entropy," *Pattern Recognit.*, vol. 56, pp. 1–15, 2016.
[204] W. Tang and K. Z. Mao, "Feature selection algorithm for mixed data with both nominal and continuous features," *Pattern Recognit. Lett.*, vol. 28, no. 5, pp. 563–571, 2007.
[205] L. Parsons, E. Haque, and H. Liu, "Subspace clustering for high dimensional data: A review," *SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 90–105, 2004.
[206] S. Goebl, X. He, C. Plant, and C. Böhm, "Finding the optimal subspace for clustering," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2014, pp. 130–139.
[207] N. Bansal, A. Blum, and S. Chawla, "Correlation clustering," *Mach. Learn.*, vol. 56, nos. 1–3, pp. 89–113, Jun. 2004.
[208] E. Rendón *et al.*, "A comparison of internal and external cluster validation indexes," in *Proc. Amer. Conf. Appl. Math. 5th WSEAS Int. Conf. Comput. Eng. Appl. (AMERICAN-MATH/CEA)*, 2011, pp. 158–163.
[209] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu, "Understanding of internal clustering validation measures," in *Proc. IEEE 10th Int. Conf. Data Mining (ICDM)*, Dec. 2010, pp. 911–916.
[210] S. Mohanavalli and S. M. Jaisakthi, "A precise distance metric for mixed data clustering using chi-square statistics," *Res. J. Appl. Sci., Eng. Technol.*, vol. 10, no. 12, pp. 1441–1444, 2015.

AMIR AHMAD received the Ph.D. degree in computer science from The University of Manchester, U.K. He is currently an Assistant Professor with the College of Information Technology, United Arab Emirates University, Al Ain, UAE. His research areas include machine learning, data mining, and nanotechnology.

SHEHROZ S. KHAN received the master's degree from the National University of Ireland Galway, Ireland, and the Ph.D. degree in computer science with specialization in machine learning from the University of Waterloo, Canada. He is currently a Scientist with the Toronto Rehabilitation Institute, Canada. He is also a Postgraduate Affiliate with the Vector Institute, Toronto. His main research interests include the development of machine learning and deep learning algorithms within the realms of aging, rehabilitation, and intelligent assistive living.