International Journal of Pure and Applied Mathematics
Volume 118 No. 7 2018, 547-556
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

ijpam.eu

# Clustering Algorithms for Mixed Datasets: A Review

K. Balaji* and K. Lavanya
School of Computer Science and Engineering
VIT University, Vellore, India

***Abstract*—Clustering is an essential technique in Data Mining which has been applied effectively in numerous perspectives. However, most of the clustering algorithms developed have been focused either on numeric or categorical datasets, but limited to both. Clustering algorithms with mixed datasets provide distance measures for handling both numeric and categorical data attributes. In this review, we present a critical analysis of the most effective algorithms described in the field of clustering algorithms with mixed datasets. Moreover, we present a comparison of these algorithms, regarding a set of features which are desirable from a practical point of view. Finally, some research lines that need to be further developed in the context of clustering mixed datasets are discussed.**

***Keywords*—Data Mining; Clustering; Mixed Data; Dissimilarity Measure;***

## I. Introduction

Clustering is an essential method in Data Mining and Pattern Recognition. This method aims at organizing a collection of objects into classes or clusters, such that objects belonging to the same cluster are similar enough to infer they are of the same type; and objects belonging to different clusters are dissimilar enough to infer they are of distinct type. There are several areas in which clustering algorithms have been successfully applied, like for instance: privacy preserving [1], information retrieval [2], text analysis [3], image processing, customer segmentation, and gene expression analysis [4].

Clustering algorithms can be classified according to different criteria, like for example, a) Clustering numeric data attributes, b) Clustering categorical data attributes, and c) Clustering mixed numeric and categorical data attributes. This work focuses its analysis on the last one criterion. Regarding to this criterion, a clustering algorithm can be classified as those algorithm that build the clusters based on the similarity relations of both numeric and categorical data attributes. Most of the clustering algorithms developed so far are based on the similarity relations either on numeric or categorical data attributes and they leave clustering with mixed datasets; these algorithms could not be suitable for applications such as heart disease [5].

The aim of this work is to present a critical review of the most influential clustering algorithms with mixed datasets reported in the literature, describing their main features and highlighting their limitations since a practical point of view.

The main objective of this review paper can be summarized as follows:

- To review the existing algorithms for discovering how concept formation is made.
- Toidentify the significances of mixed datasets clustering algorithms
- To identify the recent advances in this domain.
- To present the impacts of mixed datasets clustering algorithms in real-time applications.

The rest of this paper is as follows: In Section II the most important clustering algorithms for mixed datasets of the state of the art are described. The comparison characteristics of clustering algorithms are described in Section III. Section IV summarizes some research gaps that need to be explored during the review, and finally, Section V concludes the paper.

## II. Clusteing Algorithms for Mixed Datasets

Let X is a collection of objects. Each object is described by a set of attributes which can be numeric, categorical data attributes or both. The main aim of the clustering algorithm is to cluster the collection of objects with their mixed datasets. The existing clustering algorithms generally follow one of the following approaches:

- Conversion of categorical data attributes into numeric and performs numeric data clustering.
- Conversion of numeric data attributes into categorical and performs categorical data clustering.
- Directly handling mixed data clustering.

*Corresponding Author:
E-mail address: balaji.2016@vitstudent.ac.in

*A. Partition-based Algorithms*

The algorithms described in this subsection build the clusters based on partition. In Partition-based clustering, the center of data point becomes center of the consistent cluster. Objects are divided and clusters are updated depending on the partition. These methods are beneficial for the application of bioinformatics.

K-Prototypes algorithm [6] is a partition-based clustering algorithm which combines both K-means [7] and K-modes [8] for handling mixed datasets. The algorithm depends on the concept of K-means algorithm and furthermore, it eliminates the numeric data constraint of that algorithm. The data objects are grouped against K-Prototype and dynamically change the K-Prototype so that to get most out of within-cluster similarity of objects. Initially, K objects are selected for K clusters from the given dataset. Each object is assigned to a cluster whose model is similar according to the dissimilarity measure. After allocation of each object, the cluster model is updated. Recalculation of objects similarity in each cluster must be performed. If any object is similar to another cluster, then it is moved to that cluster. Repeat the calculation of objects similarity until no object has to modify their clusters.

K-Means Clustering for Mixed Datasets (KMCMD) algorithm [9] is a partition-based clustering algorithm which removes numeric data constraint of K-means algorithm and overcomes the complexity of K-Prototypes algorithm. A novel distance measure and cost function is recommended depending on the co-occurrence of data attributes. The algorithm randomly assigns a cluster number to all the objects. After that, the cluster center is calculated and assigns each object to the nearest cluster. The cluster center is recalculated every time whenever a new object is included. Assignment of objects and recalculation of cluster centers are repeated until the objects do not modify their clusters. Unlike K-Prototype algorithm, the importance of attribute is computed by discretizing the numeric attributes. The consequence of numeric or categorical attribute is evaluated from the distance measure rather than defined by the user. The algorithm will work not only for mixed datasets but also for pure categorical and numeric datasets.

K-Centers algorithm [10] is a partition-based algorithm which uses the concept of K-Prototypes algorithm for handling mixed datasets. K-Centers algorithm considers different frequencies for the attribute values during the update of cluster centers. The main disadvantage of K-Prototypes [6] and K-Modes [8] algorithm is that both will modify the cluster centers depending on the maximum frequency of attribute values. The cluster center which ignores the significant values of other attributes will degrade the accurateness of clustering outcome. A novel measure is proposed which considers different frequencies for attribute value on cluster centers. The K-Centers algorithm is implemented in two ways. If an object belongs to only one cluster, then it is called Hard K-Centers clustering. If an object belongs to several clusters, then it is called Fuzzy K-Centers clustering. First, it initializes the cluster center. Next, it will calculate the membership matrix. The algorithm updates the membership matrix and minimizes the cost function in order to get new cluster center. The new cluster center is computed repeatedly until cost function cannot be minimized further. The algorithm able to perform convex or spherical shape data attributes. It predefines a user-defined parameter and does not deal with outliers effectively. If the fuzzy parameter is larger, then the membership matrix fails to cluster the objects. It cannot guarantee a global optimum solution.

An Improved K-Prototype Algorithm [11] is a partition-based algorithm which presents an effective implementation for the categorical data attributes in mixed datasets and also considering the consequence of different attributes in the process of clustering. The algorithm introduces the idea of distribution centroid in categorical attributes for calculating the center of a cluster. The basic concept of fuzzy centroid [12] is used to represent the categorical data attributes in a hard clustering. The algorithm evaluates the consequence of attributes with the help of Huang's approach [13]. The maximum number of iterations, as well as the cluster number, is initialized at the beginning. The objective function is minimized by dividing the problem into sub-problems. The sub-problems are solved for numeric as well as categorical attributes to obtain the solution for the main problem. The algorithm will perform the best clustering not only on mixed datasets but also in pure numeric or categorical. In real world applications, the clustering of datasets needs the fuzzy scenario to get better results, whereas the algorithm contributes to getting accurate results based on a hard scenario.

K-Harmonic Means type Clustering algorithm for Mixed Datasets (KHMCMD) algorithm [14] is an extension of K-Harmonic Means (KHM) algorithm [15]. KMCMD algorithm suffers from initialization of centroid in a cluster. The result of clustering algorithm depends only on initial selection of centroid in a cluster. Random initialization of centroid in a cluster is a standard method. However, the results of the clustering algorithm are not comfortable with different initialization of centroid in a cluster. This issue is solved in KHM algorithm for numeric data attributes only. KHMCMD algorithm is used to solve the issue of initialization of centroid in a cluster for mixed datasets. Initially, the numeric data attributes are discretized to make categorical attributes. The dissimilarity measure proposed in [9] and cost function of KHM algorithm in a hard scenario are used for defining cluster center. The new dissimilarity measure computes the cost function in a fuzzy scenario for mixed datasets. Each object is allocated to the cluster until no data points to modify the cluster membership or a number of iterations reached.

*B. Hiearchical Clustering Algorithms*

Hierarchical clustering is used to build hierarchical structure that combines or divides the data objects into clusters. A tree is used to represent this hierarchy of cluster. Hierarchical clustering algorithms are divided into agglomerative clustering (bottom-up approach) and divisive clustering (top-down approach). The bottom-up approach of the clustering algorithm begins with only one cluster and iteratively combines two or more of the related clusters. The top-down approach of the clustering algorithm begins with a single cluster containing all objects and iteratively divides that

cluster into suitable sub-clusters. This process continues until a stopping principle accomplished.

Distance Hierarchy (DH) algorithm [16] is a hierarchical clustering algorithm which computes the similarity measure of categorical attributes and combines the result with numerical attributes. DH clustering algorithm combines several predictable distance calculation structures called simple matching method and binary encoding technique which transform categorical data attributes into numeric data attributes. The distance hierarchy is based on concept hierarchy [17-18], where the new distance measure is calculated by means of edge costs. The similarity measure of categorical data attribute is computed by the distance between the total edge costs of two nodes. DH algorithm encompasses concept hierarchy in which each edge is having a cost and simplifies the calculation of distance measure. The concept hierarchy structure consists of vertices and edges. The top-level vertices denote common concepts, whereas bottom-level vertices denote detailed concepts. The distance of two nodes is computed by the total edge costs between them. The pattern adjacency matrix is given as the input for DH algorithm. After that, the matrix is used for the consequent process of clustering. The DH algorithm is incorporated with an agglomerative hierarchical approach, so that the data analysts can reflect their knowledge for finding the similarity of data objects by using the construction of distance hierarchies.

Similarity-Based Agglomerative Clustering (SBAC) algorithm [19] is a hierarchical clustering algorithm. The algorithm uses a standard measure called Goodall [20] which processes numeric and categorical data in a common structure. After that, it can be appropriately combined with an agglomerative approach that builds a hierarchy. SBAC approach is made on the Unweighted Pair Group Method with Arithmetic (UPGMA) average [20]. The algorithm begins clustering process by using a distance matrix pair for the collection of data objects. The distance among a couple of data objects is the counterpart to their measures of similarity values. At any given time, the lowest pairwise dissimilarity data objects of clusters are combined into a distinct group. The distance between the new cluster and the old clusters are well-defined as the average distance between them. The computation of the dissimilarity measure is repeated until all the objects are combined in a single cluster. The termination of the cluster process outcomes in a dendrogram (or tree) where the leaf vertices will specifies different data objects and root vertices specifies a group which contains entire objects.

Two-step Method for Clustering Mixed Numeric and Categorical data (TMCM) algorithm [21] is a hierarchical clustering algorithm. The algorithm discovers the relationship between categorical data attributes on their co-occurrence values. All categorical data attributes are transformed into numeric data attributes, so that the overall data objects contains only numeric representation. It is very east to apply an existing algorithm for clustering process if all the data objects contain only numeric representation. At the primary step, Hierarchical Agglomerative Clustering [22] method is applied to group the initial data into some subgroups. The new shaped subgroups with additional options will be the input for K-means clustering [7] of next step. Instead of choosing individual data, every subgroup with added features will become the primary group for the K-means clustering. With the help of the primary group, K-means clustering process will be enhanced. This characteristic will become an added advantage to this algorithm which will reduce the significance of an outlier. The quality of the clustering process is computed by using the entropy.

### *C. Incremental Clustering Algorithms*

Non-incremental clustering algorithms stores and process all the input data pattern matrix in the memory. These algorithms generally need the complete input data being loaded into memory and as a result, the requirements of memory space will become high. In an incremental clustering algorithm, there is no need to load and process the whole input data in the memory. So, the required amount of memory space will become less. Incremental clustering algorithms consider the input data pattern which is to be processed one at a time in the memory. It is easy to add the new input data patterns into the existing clusters. Incremental clustering algorithms are appropriate for run-time environments as well as for very large databases.

Modified Adaptive Resonance Theory (M-ART) algorithm [23] is an incremental clustering algorithm. The algorithm uses M-ART network and concept hierarchy structure for handling mixed datasets. ART network [24] is an extremely standard incremental clustering algorithm with unsupervised neural network learning technique. Category-I ART deals with numeric data which is binary. Category-II ART deals with numeric data which is general [25]. Many data systems collect the mixed data attributes. But, Category-I ART and Category-II ART network methodologies do not deal with mixed datasets. The categorical data attributes are converted into binary information does not replicate the original information which will impact the quality of the clusters. M-ART network has two layers. Input layer consists of training datasets which comprise distance hierarchy groups, the threshold value, and stopping criterion. Initially, input records are assigned to input vectors. If the output vectors are similar which surpasses the threshold value, then we group those neurons into clusters. Otherwise, output neuron as new neurons. The process is repeated until the input record is empty or stopping criterion is met. Based on the concept hierarchy, each data attribute is associated with distance hierarchy using link costs representing the distance between two data attributes. Implementation of distance hierarchy can simplify the distance calculation.

Clustering Algorithm based on the methods of Variance and Entropy (CAVE) algorithm [26] is an incremental clustering algorithm. The algorithm calculates the similarity measure for numerical data attributes by variance and categorical data attributes by entropy. The number of clusters is predefined. Initially, the dissimilarity of two objects is computed and grouped into two different clusters. The dissimilarity of remaining records is calculated and put it into the appropriate clusters. The process is repeated until the records are empty. The algorithm can stop processing at any time and produces the output. The algorithm incrementally gets the updates of clusters whenever any new data attribute arrives at the cluster.

Mixed Self-Organizing Incremental Neural Network (MSOINN) algorithm [27] is an incremental clustering algorithm which automatically creates the number of clusters. The MSOINN algorithm is based on Adjusted Self-Organizing Neural Network (ASOINN) algorithm [28]. A novel distance measure estimates the categorical distance based on two learning techniques such as supervised and unsupervised. In supervised learning, if the value of two data attributes are related to each other, then the distance measure returns 0, otherwise 1. If the supervised learning technique is not presented, then the novel distance measure is computed using unsupervised learning technique. In unsupervised learning technique, the number of various data values are recorded in categorical data attributes and their existence occurrences in the datasets is considered. If the province size of the categorical attribute is 2, then the dissimilarity of two unrelated attributes will be larger than the instance having province size of 20. The data attributes are not clustered in the similar cluster, if the dissimilarity measure of categorical attributes is larger. The value of each data attributes is influenced on the entropy and their occurrences in common classification are the beneficial features in supervised learning technique. Initially, the clusters are applied to the neural network. Every iteration, the cluster will be deleted if it does not win through the learning process. The algorithm creates an offline level to produce an appropriate cluster number for a known dataset using the recommended dissimilarity measure and the modified rules. The labels of new instance are defined by calculating the nearest clusters using the network and clusters. The concurrent modifications of clusters are also performed.

*D. Model-based Clustering Algorithms*

Model-based clustering algorithm chooses a detailed model for every cluster and discovers the finest appropriate model. The model-based clustering is divided into two categories, such as neural network method and statistical learning method. The model requires user-defined parameters and it may change during the clustering process.

BI-Level Clustering of Mixed categorical and numerical data types (BILCOM) Empirical Bayesian algorithm [29] is a model-based clustering algorithm and uses categorical data attributes clustering as a model to lead the numerical data attributes clustering. This method performs a pseudo-Bayesian approach with categorical data attributes as the guide. In prior biological applications to genes, Gene Ontology annotations were the categorical data attribute and gene expression data was numerical data attribute. The model-based clustering algorithm discovers the gene expression with any arbitrary shaped clusters by comprising related information [30-33]. Data attributes that BILCOM clustering is especially helpful to exists within the area of medicine. The categorical data attributes denote the features or symptoms of patients and numerical data attributes denote the outcomes of medical investigations on patients. The medical results of patients are reflected in clustering the medical data sets by using this algorithm. An alternative essential application for this clustering algorithm is microarray gene expression data attributes which contain categorical data attributes indicating known gene function [34-36] and numerical data attributes indicating gene expression through tissues [37-39]. The BILCOM clustering algorithm implements clustering in two stages. The data attributes are taken from the biomedical datasets. The categorical datasets present semantic data on the objects, while numeric datasets present experimental outcomes. By using the method of Bayesian, it makes sense to use at the first stage as categorical attributes and second stage as numerical data attributes. Similarity measures for categorical data attributes are calculated first and numerical attributes are calculated next. The output of the first stage result is given as input to the second stage and the second stage is the output of this clustering algorithm.

AUTOCLASS algorithm [40] is a model-based clustering algorithm and is used to define the allocation of clusters for suitable classes which is inherited from the concepts of Bayesian approach. The algorithm discovers the foremost probable categorization of data objects in a group depends on the preceding allocation of each data attribute to the cluster and signifies the prior view to the user. In the first stage, the user chooses a probabilistic distribution for each data attributes in the dataset. Categorical data attributes are demonstrated with Bernoulli distribution, whereas numerical data attributes with a Gaussian distribution. The algorithm every time modifies the classification of objects in cluster. By considering the mean and variance which gives the maximum chance of detecting the object values, each object is allocated to cluster with the data attribute probability distributions. Furthermore, the algorithm iteratively examines a different number of clusters, which are not user specified. In each cluster, the values of mean and variance are modified by the algorithm. The algorithm iterates until the clusters and data values of probability distributions become a steady state.

Support Vector Machines (SVM) algorithm [41] is a model-based algorithm and is used to group data without any prior information of input classes. The algorithm is initialized by running an SVM classifier against data attributes with each input vector in the dataset arbitrarily categorized. The steps are repeated until an initial convergence occurs. After completion of the initialization step, the parameters of SVM for training the data attributes can be accessed. The lowest mislabeled data will be assigned to the label of other class. The algorithm will run again on the dataset and is assured to converge in this condition. Meanwhile, it converged formerly and now it has a smaller amount of data points to bring with mislabeling drawbacks. The method recovers on its uncertainly convergent result by retraining the SVM algorithm after every relabeling of the mislabeled input vectors. The repetition of the above method improves the clustering accuracy, at this point a degree of separable until misclassification occurs. SVM clustering algorithm affords a very effective mechanism to make a separating hyperplane bounded by the heaviest edge, using the training datasets. In spite of SVMs supervised nature, it has been useful to categorical data attributes to find groups in an unsupervised way. The method involves arbitrarily allocating objects to a pair of groups and re-calculating the separating hyperplane until object allocation and hyperplane is converged.

Earlier, SVM-Internal Clustering (typically stated as a one-class SVM) used internal features of SVM to discover a group as the smallest encompassing sphere in a set of data. The internal method to SVM clustering required heftiness and is

biased on the way to group with a spherical form in feature space. The SVM-Internal Clustering algorithm might only perceive the quite small cluster centers in most real-world applications. To overcome this problem, an External-SVM Clustering algorithm was presented that clusters data attributes with no preceding information of each data object classification. Primarily, in the dataset, each data object is arbitrarily categorized and training is given to the classifier of SVM. The scores of sensitivity and specificity will become low which is nearly 1 after initial convergence is achieved. The algorithm then develops the outcome, by iteratively relabeling the poorest misclassified data vectors.

*E. FuzzyClustering Algorithms*

Fuzzy clustering algorithms converts the discrete values of {0, 1} into continuous values in between [0, 1]. Fuzzy clustering describes the relationship among data objects more accurately.

General Fuzzy C-Means (GFCM) algorithm [42] is a fuzzy clustering algorithm based on the concept of Fuzzy C-Means (FCM) algorithm [43]. Frequency-based cluster models [44] are used to group categorical data attributes depends on the method of the simple matching algorithm. In FCM algorithm, only the numeric data attributes are divided into objects, whereas in GFCM numerical and categorical data attributes are divided into objects. The characteristics of the fuzzy p-mode model are defined as an array of p labels which have larger frequencies than the cluster of others. In the conventional algorithm which has single feature model, and simple matching model leads to inaccurate clusters. But GFCM algorithm has multiple labels at the categorical data attributes and produces accurate clustering results. Initially, the membership degree is chosen. The membership objective function should be minimized and dissimilarity measures are calculated. The clustering process is repeated until stopping criterion is met.

Kullback-Leibler Fuzzy C-Means Gaussian Mixture Models (KL-FCM-GM) algorithm [45] is a fuzzy clustering algorithm and based on Gath-Geva algorithm [46] for handling mixed datasets effectively. In existing approaches, fuzzy clustering of data is carried out by fuzzy k-prototypes algorithm which uses a different variance. On the contrary, an innovative fuzzy c-means algorithm makes use of entirely probabilistic dissimilarity functional for mixed datasets is proposed. The proposed algorithm uses a fuzzy objective function normalized by Kullback-Leibler variance facts and expressed on the origin of a set of likelihood conventions concerning the method of inherited clusters. The algorithm consists of an iterative method. The given objective function is enhanced over fuzzy membership functions, parameters, and weights of clusters.

Fuzzy K-means type algorithm [47] is a fuzzy clustering algorithm and computes the impact of attributes for numeric and categorical using probabilistic dissimilarity measure. In FCM algorithm, it is not possible to calculate the mean form categorical data attributes. Instead of mean, the mode is calculated for categorical data attributes which does not reflect the original information. A novel dissimilarity measure with the definition of cluster centroid is proposed. Initially, fuzzy partition matrix and threshold values are set. After that, the dissimilarity measures for numerical and categorical attributes are computed. The clusters are updated until the stopping criterion is met.

Fuzzy K-Prototype algorithm [48] is a fuzzy clustering algorithm and used to represent cluster prototype by combining mean and fuzzy centroid. The K-Prototypes algorithm implements hard partition which results in poor clustering of data attributes within the region of boundaries. The Fuzzy K-Prototype algorithm improves hard clustering. The data attributes are grouped into different clusters which have different degree of membership functions. Initially, the number of clusters, maximum iterations, and the threshold values are set. After that, the cluster prototype is divided into two parts. The first part uses mean for computing numerical attributes and the second part uses fuzzy centroid for computing categorical attributes. The dissimilarity measure between two objects is calculated and similar data objects are grouped into a single cluster. The clustering process is repeated until maximum iterations or stopping criterion is reached.

*F. Artificial Neural Networks Clustering Algorithms*

Artificial Neural Networks Clustering is based on the idea of competitive learning technique. It is divided into two categories such as hard competitive learning method and soft competitive learning method. In hard competitive learning, only the winning neuron is permitted for learning, whereas in soft competitive learning, all neurons in the network can get a chance for learning. The hard competitive learning method is the winner-take-all learning method, whereas the soft competitive learning method is the winner-take-competitive learning method.

Mixed-type Self-Organizing Map (MixSOM) algorithm [49] is an artificial neural network clustering which extends self-organizing map model to perform visualized analysis of mixed datasets. The prototype combines the features of Generalized SOM (GSOM) [50] and Visualized SOM (ViSOM) [51-52]. MixSOM modifies the distance hierarchy representation of GSOM into a more convenient representation of numeric and categorical data attributes. The algorithm visualizes the data in the form of high-dimensional space and estimates into two-dimensional space. The distance hierarchy method considers meaningful characteristics of categorical attributes during the training process. The algorithm reflects the association between model distance and SOM map distance through attributes during the modification process. The distance between the neighbors neurons are constrained to the way of predetermined attributes by making visualization of mixed data attributes. The structure of the cluster is controlled by user defined inputs. The dissimilarity measure of categorical data attributes is computed by distance hierarchy. Each input attributes and each item in the model is combined to their related distance hierarchies. The distance hierarchies which are distinct is computed and then combined.

Growing Mixed-type SOM (GMixSOM) algorithm [53] is an artificial neural network clustering algorithm. The algorithm uses distance hierarchy representation and develops excellence of projection map. The self-organizing map will develop from primary neurons to a large size of neurons during

training process. If the map is active, then it deals with the convenient arrangement of neurons instead of fixed size. However, the Growing SOMs is only recommended for the framework of handling numeric attributes. The categorical attributes are converted into numeric ones. But it does not reflect the original information of categorical attributes. Initially, training datasets, set of distance hierarchies, spreading factor and number of training step are given as input to the algorithm. The input neurons are initialized with randomly selected weights. After that, threshold value is defined according to the spreading factor. During growing stage, the size of neighborhood neurons and their learning rate are set. Each neuron errors are reset in training process. After that, the best matching unit of input vector is identified and their neighbors are updated. The error of best matching unit is computed. The steps are repeated until all the neurons are trained.

TABLE I. COMPARISON CHARACTERISTICS OF CONCEPTUAL CLUSTERING ALGORITHMS FOR MIXED DATASETS

| Algorithm | Scalability | Shape of Cluster | Sensitivity to Noise /Outliers | High-Dimensional Data | Input data in any order | Interpretation of Results | Depending on prior knowledge and user-defined parameter |
|---|---|---|---|---|---|---|---|
| K-prototypes | Yes | Convex | Yes | Yes | Yes | Yes | Yes |
| KMCMD | Yes | Convex | No | Yes | Yes | Yes | Yes |
| K-centers | Yes | Convex | Yes | Yes | Yes | Yes | Yes |
| ImprovedK-prototype | Yes | Convex | No | Yes | Yes | Yes | Yes |
| KHMCMD | Yes | Convex | No | Yes | Yes | Yes | Yes |
| DH | Yes | Arbitrary | No | Yes | Yes | Yes | Yes |
| SBAC | No | Arbitrary | No | No | Yes | Yes | Yes |
| TMCM | No | Arbitrary | Yes | No | Yes | Yes | Yes |
| M-ART | Yes | Arbitrary | No | No | Yes | Yes | Yes |
| CAVE | No | Arbitrary | No | No | Yes | Yes | Yes |
| MSOINN | No | Arbitrary | No | Yes | Yes | Yes | Yes |
| BILCOM | No | Arbitrary | Yes | No | No | Yes | Yes |
| AUTOCLASS | No | Arbitrary | No | No | Yes | Yes | Yes |
| SVM Clustering | Yes | Arbitrary | Yes | Yes | Yes | Yes | Yes |
| GFCM | No | Convex | Yes | No | Yes | Yes | Yes |
| KL-FCM-GM | Yes | Arbitrary | No | Yes | Yes | Yes | Yes |
| Fuzzy K-means | Yes | Arbitrary | Yes | Yes | Yes | Yes | Yes |
| Fuzzy K-prototype | Yes | Arbitrary | No | Yes | Yes | Yes | Yes |
| MixSOM | Yes | Arbitrary | No | Yes | Yes | Yes | Yes |
| GMixSOM | Yes | Arbitrary | No | Yes | Yes | Yes | Yes |
| FMSOM | Yes | Arbitrary | No | Yes | Yes | Yes | Yes |
| UFLA | Yes | Arbitrary | No | Yes | Yes | Yes | Yes |

Frequency neuron Mixed Self-Organizing Map (FMSOM) algorithm [54] is an artificial neural network clustering algorithm. The algorithm processes categorical data attributes directly without any conversion procedure. The algorithm is designed to address the issues present in existing algorithms GSOM [49], MixSOM [50], CPrSOM [55], and NCSOM [56]. FMSOM is based on the concept of NCSOM, but includes the likelihood tables from CPrSOM. FMSOM has the ability to train the neurons in an effective and precise manner. Unlike NCSOM, the algorithm converges after a finite number of steps. FMSOM constructs a novel prototype to handle categorical data attributes or mixed datasets. Initially, the dataset is given as input and number of iterations, size of the map, radius, and neighborhood degeneration are initialized.

SOM topology is created and reference vectors are initialized to random values. FMSOM algorithm consists of three phases. In competitive phase, dissimilarity of numerical attributes is calculated by using the Classic SOM algorithm such as Euclidean distance and dissimilarity of categorical attributes is calculated by using the measure of probability. In second phase, depending upon the computation of winning neuron the cooperative process begins. The Gaussian neighborhood of the winning neuron is computed and updated. In third phase, adaptation process takes place. The updates for neuron's weight vectors are calculated for mixed data attributes. The clustering process terminates after a finite number of iterations.

Unsupervised Feature Learning with Fuzzy ART (UFLA) algorithm [57] is an artificial neural network clustering algorithm. The numerical and categorical features are represented in the form of sparse features. Initially, data pre-processing is performed for missing values, interval, and multi-

TABLE I (CONTINUED)

| Algorithm | Data Structure | Type of Cluster | Representation of Cluster | Time Complexity |
|---|---|---|---|---|
| K-prototypes | Set | Dynamic | Disjoint | $O((s+1)cn)$ |
| KMCMD | Set | Dynamic | Disjoint | $O(a^2n+a^2C^3+sn(ct_n+ct_cC))$ |
| K-centers | Set | Incremental | Fuzzy | $O(n)$ |
| ImprovedK-prototype | Set | Dynamic | Disjoint | $O(c(t+t_n+St-St_n)ns)$ |
| KHMCMD | Set | Incremental | Fuzzy | $O(t^2n+m^2C^3+sn(ct_n+st_cP))$ |
| DH | Hierarchical | Incremental | Disjoint | $O(n^2)$ |
| SBAC | Hierarchical | Dynamic | Disjoint | $O(n^2)$ |
| TMCM | Hierarchical | Dynamic | Disjoint | $O(n^2)$ |
| M-ART | Hierarchical | Incremental | Disjoint | O(R * D * O * DI) |
| CAVE | Hierarchical | Incremental | Disjoint | $O(N^2)$ |
| MSOINN | Hierarchical | Incremental | Disjoint | O(N) |
| BILCOM | Set | Static | Disjoint | $O(n^2)$ |
| AUTOCLASS | Set | Static | Disjoint | $O(cd^2ns)$ |
| SVM Clustering | Set | Static | Disjoint | $O(n)$ |
| GFCM | Hierarchical | Dynamic | Fuzzy | $O(n)$ |
| KL-FCM-GM | Hierarchical | Dynamic | Fuzzy | $O(n)$ |
| Fuzzy K-means | Hierarchical | Dynamic | Fuzzy | $O(s(tn+S^2cs+nc+nct_n+nct_cS))$ |
| Fuzzy K-prototype | Hierarchical | Dynamic | Fuzzy | $O(t^2n+t^2S^3+c(t+s+St-St_n)ns)$ |
| MixSOM | Hierarchical | Dynamic | Fuzzy | Type + Layer |
| GMixSOM | Hierarchical | Dynamic | Fuzzy | Type + Layer |
| FMSOM | Hierarchical | Dynamic | Fuzzy | Type + Layer |
| UFLA | Hierarchical | Dynamic | Fuzzy | Type + Layer |

s – No. of iterations; c – No. of clusters; n – No. of objects; a – Total no. of attributes; C – Average no. of distinct categorical values; P – No. of attribute values of categorical attributes; t – total no. of attributes; $t_n$ – No. of numerical attributes; $t_c$ – No. of categorical attributes; S – Maximal no. of values for categorical attributes; R – Training Round; D – Training data record; DI – Data Dimension; O – Output Neuron Number; N – Dataset Size; d – dimensionality;

value data. Binary values are assigned to categorical data and normalizing the numeric data. The number of clusters is predetermined. Weights are defined for each cluster. The algorithm clusters entire dataset to produce the model and passed to the feature encoder. The feature encoder will encode mixed data type into sparse representation. After that, the clustering will be performed by traditional clustering algorithm. The distance measure is calculated and clusters are updated until the stopping criterion is met.

## III. Comparision characteristics of Clustering Algorithms for Mixed Datasets

Table I shows the comparison characteristics of clustering algorithms described in Section II.

## IV. Discussion-Gaps

Several studies around clustering algorithms obtain consideration of the researchers. Even though, several realizations have been accomplished, yet, there are still missing gaps that need to be filled. We review about the gaps in conceptual clustering algorithms, as follows:

- Many clustering algorithms are not able to handle mixed data attributes directly which are necessary for today's real-time applications. Many algorithms transform one type of attribute into other, which produce loss of information. We need to develop an efficient and accurate clustering algorithm that able to handle both mixed attributes and missing data.
- Most of the algorithms produce disjoint clusters. But some algorithms that producefuzzy clusters with different membership degree provide a good interpretation of clusters. We need to develop research in building fuzzy clusters and generate concept description that can accurately produce results.
- The hierarchical representation of clusters is computationally expensive. But hierarchical structure provides necessary information for the user. We need to develop an efficient clustering algorithm able to build hierarchical of clusters and concepts.
- Most of the algorithms result in poor clustering due to addition, deletion, and modification of objects during run time. We need to develop efficient and accurate clustering algorithm that able to process additions, deletions, and modifications of objects.
- Most of the fuzzy clustering algorithms provide good clustering. We need to develop efficient fuzzy clustering algorithms with the characteristics such as dynamic and easy interpretation of clusters.

## V. Conclusion

One of the most essential properties of the clustering algorithm is to handle mixed datasets effectively. In this review paper, clustering algorithms for mixed datasets are discussed with their limitations. After that, comparison characteristics of all algorithms are presented. Finally, some research gaps that need to further explore are discussed.

## References


[1] M.Z. Islam, L. Brankovic, Privacy preserving data mining, a noise addition framework using a novel clustering technique. Knowledge-Based Systems, 2011. 24(8):p. 1214-1223. http://dx.doi.org/10.1016/j.knosys.2011.05.011

[2] G. Bordogna, G. Pasi, A quality driven Hierarchical Data Driven Soft Clustering for information retrieval. Knowledge-Based Systems, 2012. 26(1):p. 9-19. http://dx.doi.org/10.1016/j.knosys.2011.06.012

[3] W. Zhang, T.Yoshida, X.J. Tang, Q. Wang, Text clustering using frequent itemsets. Knowledge-Based Systems, 2011. 23(5):p. 379-388. http://dx.doi.org/10.1016/j.knosys.2010.01.011

[4] W. Chen, G. Feng, Spectral Clustering, a semi-supervised approach. Neuro-computing, 2012. 77(1):p. 229-242.

[5] Jiawei Han and Micheline Kamber, Data Mining Concepts and Techniques. Morgan Kaufmann Publishers, 2006.

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

[7] MacQueen J, Some methods for classification and analyis of multivariate observations. Proc Fifth Berkeley Symp Math Stat Probab, 1967. 1:p. 281-297.

[8] Z. X. Huang, Extensions to the K-means algorithm for clustering large datasets with categorical values. Data Mining and Knowledge Discovery, 1998. 2(3):p. 283-304.

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

[10] Wei-Dong Zhao, Wei-Hui Dai, and Chun-Bin Tang, K-Centers algorithm for clustering mixed type data. PAKDD, Springer-Verlag Berlin Heidelberg, 2007.

[11] J. Ji, T. Bai, C. Zhou, et al., An improved k-prototypes clusteirng algorithm for mixed numeric and categorical data. Neuro computing, 2013. 120:p. 590-596.

[12] W. Kim, K. H. Lee, D. Lee, Fuzzy clustering of categorical data using fuzzy centroids. Pattern Recognition Letters, 2004. 25(11):p. 1263-1271.

[13] Z. X. Huang, M. K. Ng, H. Q. Rong, et al, Automated variable weighting in k-means type clustering. IEEE Transaction Pattern Analysis Machine Intelligence, 2005. 27(5):p. 657-668.

[14] Amir Ahmad, Sarosh Hashmi, K-Harmonic means type clustering algorithm for mixed datasets. Applied Soft Computing, 2016. http://dx.doi.org/10.1016/j.asoc.2016.06.019

[15] B. Zhang, Generalized K-Harmonic Means. Hewlett-Packard Laboratoris Technical Report, 2000.

[16] Chung-Chian Hsu, Chin-Long Chen, Yu-Wei Su, Hierarchical clustering of mixed data based on distance hierarchy. Information Sciences, 2007. 177:p. 4474-4492. http://dx.doi.org/10.1016/j.ins.2007.05.003

[17] J. Han, Y. Fu, Dynamic generation and refinement of concept hierarchies for knowledge discovery in databases. In: Proceedings of the AAAI'94 Workshop Knowledge Discovery in Databases (KDD'94), Seattle, 1994.

[18] J. Han, Y. Cai, N. Cercone, Data-driven discovery of quantitative rules in relational databases. IEEE Transactions Knowledge on Data Engineering, 1993. 5:p. 29-40.

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

[20] D. W. Goodall, A New Similarity Index Based On Probability. Biometrics, 1966. 22:p. 882-907.

[21] Ming-Yi Shih, Jar-Wen Jheng and Lien-Fu Lai, A Two-Step Method for Clustering Mixed Categorical and Numeric Data. Tamkang Journal of Science and Engineering, 2010. 13(1):p. 11-19.

[22] C. Hsu, Y. P. Huang, Incremental Clustering of mixed data based on distance hierarchy. Expert System Applications, 2008. 35(3):p. 1177-1185.

[23] Carpenter G, Grossberg A S, Rosen D B, Fuzzy ART, Fast stable learning and categorization of analog patterns by an adaptive resonance system. Neural Networks, 1991. 4:p. 759-771.

[24] Carpenter G, Grossberg A S, ART2, Self-organization of stable category recognition codes for analog input patterns. Applied Optics: Special Issue on Neural Networks, 1987. 26:p.4919-4930.

[25] C. C. Hsu, Y. C. Chen, Mining of mixed data with application to catalog marketing. Expert System Applications, 2007. 32(1):p. 12-23.

[26] Fakhroddin Noorbehabahani, Sayyed Rasoul Mousavi, Abdolreza Mirazaei, An Incremental mixed data clustering method using a new distance measure. Soft Computing, 2015. 19:p. 731-743.

[27] Shen F, Hasegawa O, A fast nearest neighbor classifier based on self-organizing incremental neural network. Neural Networks, 2008. 21(10):p. 1537-1547.

[28] B. Andreopoulos, A . An and X. Wang, Bi-level clustering of mixed categorical and numerical biomedical data. International Journal of Data Mining and Bioinformatics, 2006. 1(1):p. 19-56.

[29] B. Adryan and R. Schuh, Gene ontology-based clustering of gene expression data. Bioinformatics, 2004. 20(16):p. 2851-2852.

[30] R. Bellazzi and B. Zupan, Towards knowledge-based gene expression data mining. Journal of Biomedical Informatics, 2007. 40(6):p. 787-802.

[31] M. Brown, W. Grundy, D. Lin, et al, Knowledge-based analysis of microarray gene expression data by using support vector machines. PNAS, 2000. 97(1):p. 262-267.

[32] C. Pasquier, F. Girardot, K. Jevardat de Fombelle, et al, THEA, Ontology driven analysis of microarray data. Bioinformatics, 2004. 20:p. 2636-2643.

[33] Dwight S S, Harris M A, Dolinski K, et al, Saccharomyces Genome Database provides secondary gene annotation using the gene ontology. Nucleic Acids Research, 1999. 30:p. 69-72.

[34] Gene Ontology Consortium, Creating the gene ontology resource: design and implementation. Genome Research, 2001. 11:p. 1425-1433.

[35] Lord P W, Stevens R D, Brass A, et al, Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. Bioinformatics, 2003. 19:p. 1275-1283.

[36] Eisen M B and Brown P O, DNA arrays for analysis of gene expression. Methods Enzymol, 1999. 303:p. 179-205.

[37] Eisen M B, Spellmen P T, Brown P O, et al, Cluster analysis and display of genome-wide expression patterns. Proc.Natl.Acad.Sci., USA, 1998. 95(25):p. 14863-14868.

[38] Slonim D K, Tamayo P, Mesirov J P, et al, Class prediction and discovery using gene expression data. In: Proceedings of 4th International conference on Computational molecular biology, Tokyo, Japan, 2000.

[39] Stuz J, and Cheeseman P, Bayesian classification (AUTOCLASS), theory and results. Advances in Knowledge Discovery and Data Mining, 1995.

[40] S. Winters-Hilt and S. Merat, SVM Clustering. BMC Bioinformatics, 2007.

[41] Mahnhoon Lee, Witold Pedrycs, The fuzzy C-means algorithm with fuzzy P-mode prototypes for clustering objects having mixed features. Fuzzy Sets and Systems, 2009.

[42] J C Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York, 1981.

[43] D W Kim, K H Lee, D Lee, Fuzzy clustering of categorical data using fuzzy centroids. Pattern Recognition Letters, 2004. 25:p. 1263-1271.

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

[45] Gath I, Geva A B, Unsupervised optimal fuzzy clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1989. 11(7):p. 773-781.

[46] Sonbaty, Yaseer Ei, M A Ismail, Fuzzy clustering for Symbolic data. IEEE Transaction on Fuzzy Systems, 1998.

[47] J Ji, W Pang, C Zhou, et al, A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data. Knowledge-Based Systems, 2012. 30:p. 129-135.

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

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

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

[51] H Yin, Data visualization and manifold mapping using the ViSOM. Neural Networks, 2002. 15(9):p. 1005-1016.

[52] Wei-Shen Tai, Chung-Chian Hsu, Growing Self-Organizing Map with cross insert for mixed-type data clustering. Applied Soft Computing, 2012. 12:p. 2856-2866.

[53] Carmelo del Coso, Diego Fustes, Carlos Dafonte, et al, Mixing numerical and categorical data in a Self-Organizing Map by means of frequency neurons. Applied Soft Computing, 2015. 36:p. 246-254.

[54] M. Lebbah, K Benabdeslm, Visualization and clustering of categorical data with probabilistic self-organizing map. Neural Comput. Applications, 2010. 19:p. 393-404.

[55] N Chen, N C Marques, An extension of self-organizing maps to categorical data. In: Proceedings of the 12th Protuguese conference on Progressin Artificial Intelligence, EPIA, 2005.

[56] Dao Lam, Mingzhen Wei and Donald Wunsch, Clustering Data of Mixed Categorical and Numerical Type with Unsupervised Feature Learning. IEEE Transactions, 2015.