ISSN: 0976-3104
SPECIAL ISSUE: Emerging trends in Computer Engineering and Research (ECER)

REVIEW

A REVIEW OF CATEGORICAL DATA CLUSTERING METHODOLOGIES BASED ON RECENT STUDIES

N. Sowmiya[1], B. Valarmathi[2]*

[1]School of Information Technology and Engineering, VIT University, INDIA

[2]Department of Software and Systems Engineering, School of Information Technology and Engineering, VIT University, INDIA

ABSTRACT

In day to day activities, a very large volume of information is collected in all fields. The data mining task is necessary to handle those large amounts of data's. Clustering is the fundamental task in data mining, its main objective is to partition the dataset consists of 'p' objects into 'q' clusters. This paper presents the literature review of the clustering algorithm for categorical and binary attributes. Many algorithms were proposed in the literature for clustering categorical and binary data. The review is based on the type of methods used for clustering categorical data, evaluation criteria, datasets used, and input & output parameters. The objective of this review is to show which algorithm performs well when compared to the clustering accuracy obtained from various methods for similar datasets.

KEY WORDS
Data mining; Clustering; Categorical Data; Boolean values; K-means; Hierarchical;

Received: 28 May 2017
Accepted: 25 July 2017
Published: 24 Sept 2017

INTRODUCTION

In a real life senario a large a very large volume of information is collected in the field of medicine, academics, market basket data transactions, banking, and etc. To handle these large amount of data, data mining concept was evolved and still in the emerging area of research since 1960's. Data mining is an extraction of information from a large set of the database. Many data mining techniques are available for the extraction of knowledge. Some of the techniques include Classification, Clustering, Association Rule Mining, Prediction, and etc. Similarly, many algorithms were available for each technique. Our focus is on clustering technique. Clustering is used to group the similar objects together in one group and dissimilar objects in other group, the dataset is partitioned into 'q' clusters based on similarity or distance measures [58]. For good quality of the cluster, the inter-cluster similarity is less and intra-cluster similarity is more. Clustering is called as unsupervised learning because it does not use predefined classes or labels for clustering data.

Some of the requirements of clustering include scalability, ability to handle different types of data, noisy data, high dimensional data, and insensitive to the order of the input. The type of data used for the clustering algorithm includes Interval-scaled, Binary, Categorical, Ratio scaled and Attributes of mixed data types. This paper can deal with the clustering algorithm for categorical and binary data only.

There are five methods of clustering algorithms like hierarchical, partitioning, density, grid, and model-based clustering. [Fig. 1] shows the block diagram representation of the clustering methods.

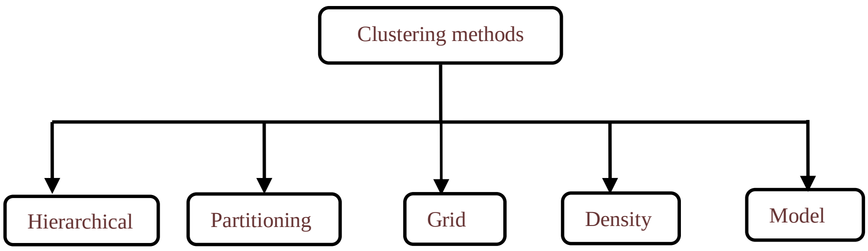

Fig. 1: Methods of clustering

The brief explanation of each of the method is discussed below

***Corresponding Author**
Email: valarmathi.b@vit.ac.in
Tel.: +91 9442811963

Hierarchical clustering algorithm groups the instances into a hierarchy or tree of clusters based on the distance measure as the criterion function. At the highest level, all items belong to the same cluster. Hierarchical clustering methods are of two types, the first approach is agglomerative (AGENS) or bottom-up approach. It starts with a single instance and successively combines the instances which are similar to one another, till all the clusters are merged into a single cluster or a stopping condition satisfies. The second approach is divisive (DIANA) or top-down approach. It starts with all the instances in one cluster. During every iteration, a cluster is split into smaller clusters, until all instances are in one cluster or a stopping condition satisfies [45].The tree representation of the hierarchical clustering was viewed by dendro gram.

THE IIOAB JOURNAL
COMPUTER SCIENCE: Guest editor-Prof. S. Prabu & Prof. Swarnalatha P

IIOAB JOURNAL
ISSN: 0976-3104

The important feature in the hierarchical clustering is that it will not assume the number of clusters. Examples of this type of clustering include ROCK, CHAMELEON, and etc.

In partitioning based clustering method, it divides a database D into 'x' partitions, where each partition corresponds to a cluster. Some examples of partitioning methods are K-means, K-medoids, and CLARANS [45].

Clusters are produced based on the density of points in density based clustering. A region with more compactness of points shows the presence of clusters, whereas regions with a low compactness of points represent the noise or outliers. DBSCAN, OPTICS, Den Clue are examples for density-based clustering [45].

In grid-based clustering, it divides the dataset consists of 'p' objects into a predetermined number of cells that form a grid structure. Few examples are STING, Wave Cluster, and CLIQUE [45].

EM, SOM, and COBWEB are the examples for model-based clustering method, in which a model is hypothesized for each of the clusters and tries to find the best fit of that model to each other in model-based clustering [45].

Some of the applications of clustering include pattern recognition, spatial data mining, World Wide Web, document clustering, image processing, cellular manufacturing, and etc.

This paper presents the literature review of the clustering algorithm for categorical and binary data, based on the type of methods used for clustering categorical data, evaluation criteria, datasets used, and input & output parameters.

[Table 1] represents the example for categorical datasets consists of nine objects, and five attributes belong to three classes. The attribute values are represented using X, Y, Z, P, Q and the corresponding attribute numbers. The whole dataset is divided into three classes namely 1, 2 and 3. The objects obj1, obj4, obj5belong to class 1, the objects obj2, obj3, obj7 belong to class 2, and similarly, the objects obj6, obj8, obj9 belong to class 3.

Table 1: A sample categorical dataset

Data Objects	$Attribute_1$	$Attribute_2$	$Attribute_3$	$Attribute_4$	$Attribute_5$	Classes
obj_1	X_1	Y_2	Z_3	P_1	Q_3	1
obj_2	X_1	Y_2	Z_3	P_1	Q_3	2
obj_3	X_2	Y_2	Z_3	P_2	Q_3	2
obj_4	X_2	Y_1	Z_1	P_2	Q_2	1
obj_5	X_1	Y_2	Z_1	P_4	Q_2	1
obj_6	X_3	Y_3	Z_1	P_4	Q_1	3
obj_7	X_2	Y_3	Z_2	P_4	Q_1	2
obj_8	X_3	Y_2	Z_2	P_3	Q_1	3
obj_9	X_3	Y_3	Z_2	P_3	Q_2	3

Steps involved in clustering

The objective of clustering is to combine the most similar objects into groups. The steps involved in the clustering are given as follows:

Step 1: Find the similarity/dissimilarity of the data objects using the distance measures
Step 2: Find the method used to form the clusters
Step 3: Decide the input parameters (e.g. Number of clusters)
Step 4: Decide the output parameters (e.g. Cluster validation measure)

Similarity and dissimilarity measures

Distance metrics are generally used to find the similarity and dissimilarity of the objects. Many benchmark distance metrics are available. Some of the commonly used distance metrics are given in [Table 2].

Table 2: Commonly used distance metrics

S.No.	Distance Measure
1.	Chebychev
2.	City block
3.	Correlation
4.	Cosine
5.	Euclidean
6.	Hamming

IIOAB JOURNAL ISSN: 0976-3104

7.	Jaccard
8.	Mahalanobis
9.	Manhattan
10.	Minkowski
11.	Seuclidean
12.	Spearman

Other than the benchmark distance measures mentioned in [Table 1], many researchers have developed their own similarity measures. “Some of the available similarity indexes present in literature are Gower similarity (GOW) (Gower, 1971), Eskin similarity (ESK) (Eskin, Arnold, Prerau, Portmoy & Stolfo, 2002), Inverse Occurrence Frequency similarity (IOF) (Church & Gale, 1995), Occurrence Frequency similarity (OF), Good all similarity (GOO), Gambaryan similarity (GAM), Lin similarity (LIN), Anderberg similarity (AND), and Smirnov similarity (SMI). These are the similarity measures compared based on the four validation measure viz., NCC, compactness, entropy and silhouette index for 15 datasets by [35]. Whereas, [75] proposed a similarity measure MCSM (Multiple Categorical Similarity Measure) for multiple categorical datasets”.

Methods for clustering data

There are five methods of clustering algorithms available in the literature “hierarchical clustering method, partition based method, density based method, grid based method, and model-based clustering method”. The detailed review of the methods is given in section 3 of this paper.

Input parameters

In some clustering algorithms, the input parameter ‘k’ which is nothing but the number of clusters should be known before performing the clustering. Example for this type of clustering is partition based clustering methods like k-means, k-medoids, k-modes, and etc. The number of given classes are three in [Table 1]. While performing clustering the number of clusters should be given as a three. Instead, if the number of clusters is given as a two means the clustering accuracy varies automatically. So the clustering accuracy depends on the number of clusters, we should have a prior knowledge about the required number of clusters. [24] proposed an algorithm called categorical data clustering with subjective factors (CDCS). The main feature of the algorithm is automatically decides the proper clustering parameters. “[26] developed a categorical data clustering method named BK Plotto validate the clusters. (Kuo et al. 2014) developed an automatic clustering algorithm called automatic kernel clustering with bee colony optimization (AKC-BCO). It automatically decides the number of clusters and assigns data points to correct clusters”.

Validation Measures

The validation measures compute the performance of the clustering algorithm. It is defined by combining compactness and separability [82]. Compactness used to measure the closeness of the cluster objects. Separability is used to measure the distinctness between the clusters. The types of validation measures available are internal validation measure and external validation measure. The first method used to evaluate the goodness of the clusters without any external information and the second method used to evaluate the clustering results by comparing the results with the externally supported class labels. The validation measures used in various studies are given in section 3.3.

Some of the commonly used internal measures are Davies-Bouldin index, Silhouette index, Bayesian information criterion (BIC), Dunn Index, etc. And the most widely used external measures are Normalized Mutual Index (NMI), Purity or Rand Index (RI), Entropy, F-measure, Adjusted Rand Index (ARI).

REVIEW METHODOLOGY

This paper presents the review based on the type of data used for clustering based on similarity or dissimilarity measure used in each article, the methods/techniques used to form clusters, the number of datasets, validation measures and tools & the system specifications of the model developed in each reference.

Methods of clustering

Cluster analysis can deal with four types of data, Interval-scaled, binary, categorical, and ratio scaled values. In this study, the focus is only on the binary and categorical data types.

IIOAB JOURNAL ISSN: 0976-3104

Table 3: Data types incorporated in the articles

Data type	Number of articles	Articles
Binary	8	[16], [42], [43], [69], [89], [90],[96],[97]
Categorical	67	[3], [4],[6], [9],[10], [13], [11], [12], [14],[17], [18], [19],[20],[22],[21], [23], [24], [25], [26], [27],[28], [29],[31], [33], [34], [36],[37],[38], [40],[41],[44],[48],[49], [50], [51], [52],[54], [55],[56],[57], [62], [63], [64], [66], [67],[68], [70], [71], [72], [73], [74], [77],[78], [79], [80],[81], [85],[86], [87], [88],[91], [92], [93], [95], [98], [100]
Mixed numeric and categorical data	11	[1], [2], [15], [30], [32], [53], [59], [60], [61],[75], [99]

[Table 3] shows the details of the articles which have used the pure categorical data, pure binary data, and mixed numeric & categorical data types. The detailed review of the clustering algorithms with respect to the basic methods or division of clustering is discussed in this section.

Partition based method

Many algorithms were available in the literature for clustering larger datasets. The algorithms like "CLARANS proposed by Ng and Han (1994), BRICH by [100], and DBSCAN by (Ester et al. 1996) are suitable for solving numerical datasets" only and not applicable for solving categorical dataset [55].

Ralambondrainy (1995) proposed a categorical clustering algorithm using k-means algorithm by converting the categorical values into binary values. This approach treats the binary values as numeric values and performs "k-means clustering". The disadvantage of this algorithm includes the "computational cost" and the mean values between 0 and 1 do not signify the uniqueness of the clusters.

[55] proposed two algorithms for clustering categorical data by extending the k-means algorithm. First one is k-modes algorithm by replacing mean by mode, it used a simple matching distance measure for clustering categorical attributes. In order to reduce the computational cost a "frequency based method" was used to recalculate the modes. Second is "k-prototype algorithm" by combining "k-means and k-modes algorithm" for clustering data with mixed numeric and categorical attributes.

Many partitioning based clustering algorithms required a random selection or pre-setting of initial points (mean or modes) of the clusters for clustering. Choosing of these initial points randomly will leads to different cluster results. So, [88] did an experimental study on the refinement of initial points to "k-modes" type categorical clustering algorithm for the better clustering results. Based on the experimental study they found that, k-populations algorithm produced better clustering results.

[75] mentioned that the performance of k-modes, k-prototypes and fuzzy k-modes algorithms results in local optimum only. So, a tabu search method for obtaining the global optimum results for categorical data is proposed.

[63] developed a new fuzzy based clustering method by extending "fuzzy k-modes" algorithm for clustering categorical data. In that, the hard-type centroids were replaced by fuzzy centroids in order to fully exploit the power of fuzzy sets. The proposed method was compared with the two existing algorithms namely "k-modes and fuzzy k-modes" and reported that it produced better clustering results.

[2] developed a "k-means" type model for categorical and numeric data clustering. The modified description of the initial points was introduced to conquer the numeric data alone constraint of the traditional "k-means" algorithm. A novel cost function and a dissimilarity measure were also proposed. The proposed algorithm was tested on real life datasets.

"[9] developed a clustering algorithm for handling high-dimensional categorical data by extending the "k-modes" algorithm using optimization methods". [9, 10, 11] experimented a k-mode type algorithm which automatically initializes the cluster centers and the number of clusters. Similarly, (He et al.2011) useda "k-modes algorithm using attribute value weighting" in the distance computation.

(Hatamlou 2012) introduced a new partitioning based algorithm using the concept of binary search algorithm. The initial centroids were chosen from the different parts of the dataset. It is noted that it converged to the same results in different runs. [15] proposed a geometric codification for clustering mixed categorical and numeric data. It codified the categorical attributes into numerical values and performed numerical clustering algorithm by combined with k-means algorithm.

[21] developed a new distance measure and a rough membership function to overcome the limitation of simple matching distance measure and Ng's distance measure for the k-modes algorithm for clustering categorical data. "[12] proposed a "weighting k-modes algorithm" for categorical data to perform subspace

clustering". In addition to the usual k-modes clustering procedure, a step to calculate weights automatically using complement entropy for all the dimensions in each and every cluster was added.

[60] developed another version of "k-prototype algorithm" for clustering "numeric and categorical data". To represent a prototype of clusters the mean & fuzzy centroids were combined and in order to calculate the distance among instances and the prototypes a distance measure was developed. "Similarly, (Ji et al. 2013) developed an improved k-prototype algorithm for mixed numeric and categorical data", here the prototype of the "categorical attributes in the cluster was represented by distribution centroids and to represent the prototype of a numerical attributes in a cluster the mean and the distribution centroids". A new dissimilarity measure was proposed to find the distance between the instances and the prototypes. In both methods, the performance of the algorithm was tested for four real world datasets and the results were compared with the traditional clustering algorithms.

"[86] proposed a medoids based clustering method called k-Approximate Modal Haplotype (k-AMH). k-AMH is a medoids based clustering for clustering categorical data and it was compared with the centroids based clustering methods like k-modes, k-population, and fuzzy k-modes algorithm in terms of clustering accuracy. [87] enhanced the k-AMH algorithm using the same procedure as that of k-AMH, termed as (Nk-AMH I), (Nk-AMH II),and (Nk-AMH III) but with the addition of two methods likely new initial center selection and new dominant weighting methods for clustering categorical data based on optimization and fuzzy procedures.

[11] proposed a fuzzy clustering algorithm by modifying the objective function of the fuzzy k-modes algorithm by including between cluster information to minimize the within cluster dispersion and between cluster partition simultaneously. [7] proposed a k-modes type clustering algorithm for categorical data. The objective function is modified by adding the between cluster similarity term in it, to overcome the limitation of weak separation of clusters in usual clustering algorithms. The algorithm was tested for some real world datasets and reported that this method produced better results than original counterparts in categorical data clustering and applicable for large datasets.

[92] compared the performance of the objective functions of the algorithms like k-medoids, k-modes, and within cluster dispersion analytically. Also, they verified the objectives for real valued datasets. The experiments were conducted to prove the performance of the objective function using the real-life data sets and reported that within cluster dispersion algorithm performs better than other methods two methods. Similarly, [8] compared the generalization, effectiveness and normalization objective functions of the internal validity functions like k-modes, category utility function, and the information entropy function by using the developed generalized validity function for evaluating the categorical data results in a solution space. Also, they addressed the problem while using these validity functions for evaluating the clusters whether the between cluster information is ignored".

Hierarchical based methods

"ROCK, a robust clustering algorithm for categorical and binary data using links was proposed by (Guha et al.2000)". It overcomes the drawbacks of the traditional clustering algorithms using distance measure or similarity measure. Using distance measure or similarity measure for clustering categorical and binary data is not appropriate. So, the concept of 'link' was introduced to find the common neighbors between the data points. The performance of the algorithm was tested on three datasets like, mushroom, congressional votes, US Mutual funds. "A quick version of the ROCK algorithm called QROCK" was proposed by [36] based on the concept of graphs. The final clusters were the components of the graph and the data points as the vertices. The main advantage of the QROCK over the ROCK algorithm was, the computation time of QROCK was reduced because of the 'merge' and 'find' concept introduced in the ROCK algorithm.

"[89] proposed a hierarchical clustering algorithm for binary gene expression data". [5] developed a scalable clustering algorithm called LIMBO, a bottleneck information framework for the design of the novel distance measure for the categorical attributes was used. It is a kind of hierarchical clustering algorithm. The main advantage of the LIMBO was in single execution and it could produce the clusterings of different sizes.

(Barbara et al. 2002) proposed COOLCAT, clustering algorithms for the categorical data based on entropy. It is applicable for both categorical data and also data streams. Entropy is lower for clusters having similar objects and it is higher for clusters having dissimilar objects.

"[57] proposed a framework to learn a context-based dissimilarity measure for categorical attributes". Based on the distribution of objects in other attributes, the distance between two objects of an attribute is determined. This method is embedded in hierarchical clustering method to validate the proposed method.

[93] developed a divisive hierarchical clustering termed as DHCC. The task of categorical data clustering was viewed in the type of optimization point of view and proposed a procedure for initialization and splitting of clusters. The advantages of this method is, it performs automatic clustering, "the dendro gram representation is obtained due to the hierarchical nature of the algorithm, the order of the data is independent, scalable for large dataset, and finding clusters in subspaces".

IIOAB JOURNAL ISSN: 0976-3104

"[80] proposed an information theory based hierarchical divisive clustering algorithm for categorical data using the mean gain ratio (MGR) of the attributes. The attribute having highest MGR is selected as the clustering attributes and equivalence class with minimum entropy is determined as the cluster and the other equivalence class is considered as the new dataset and repeats the process until all the instances are grouped into the clusters. The performance of the MGR was compared with the existing other four algorithms based on the entropy or mutual information such as COOLCAT (Barbara et al. 2002), MMR [79], K-ANMI [51], G-ANMI [33]".

Density-based clustering methods

[4] enters proposed a hierarchical density-based clustering method for categorical data named as HIERDENC and also developed a complementary index for searching dense subspaces. In that, the data was represented in the form of cube, where there is no ordering of the instances. Because of this advantage if the new instance enters into the system the HIERDENC index is only updated and the re clustering was not required. Initially the formation of clusters was started from the dense regions of the cube. Later the close by dense regions was connected to form further clusters. The HIERDENC method was compared with few existing categorical clustering algorithms and reported that the algorithm performed better scalability, runtime and cluster quality on large datasets.

[13] proposed an enhanced DBSCAN algorithm for incrementally building and updating of arbitrarily shaped clusters in large datasets. Instead of searching the whole dataset it searches only the partitions, this leads to the betterment of the results when compared with the other incremental clustering algorithms.

Model-based methods

One-dimensional Clustering is nothing but clustering data by considering all the attributes in the dataset. This way of clustering is not appropriate for complex datasets with many attributes. To overcome this draw back "[29] proposed a model based method for clustering multidimensional categorical data".

(Bauldry et al. 2015) proposed a model based algorithm which it directly finds the number of clusters and also can handle the external variables. [91] proposed a model based method based on the mixture of latent trait models with common slope parameters for clustering binary data. To determine the model parameters by means of fast algorithms the various approximations to the likelihood is exploited.

Artificial intelligence based methods

"The fuzzy k-modes algorithm is efficient for clustering categorical data. The fuzzy objective function is minimized when the algorithm searches for the fuzzy membership matrix. So, the fuzzy k-modes algorithm may stop at local optimal solution. To overcome the drawback of the fuzzy k-modes algorithm [38] proposed a genetic fuzzy k-modes algorithm for clustering categorical data. Where, the GA and fuzzy k-modes algorithms were hybridized to find the global optimal solutions. This algorithm was tested for two real life datasets and the performance was compared".

Many researches were found solutions for the categorical data clustering using the single measure for finding the clusters. This may not suitable for different datasets. To overcome this [74] developed a multi-objective genetic algorithm based fuzzy clustering for categorical data. "The two objective functions optimized by the proposed method are fuzzy compactness and the fuzzy separation of the clusters". This method was compared qualitatively and quantitatively with other algorithms and also it was tested for synthetic and real life datasets.

[66] proposed a self-organization map (SOM) for clustering and visualization of categorical data based on the Kohonen map. [53] proposed an extended SOM called MixSOM algorithm for clustering mixed numeric and categorical data.

Few authors proposed a single objective function for clustering categorical data. Such single objective function may be inappropriate for all type of datasets. So, in order to overcome this drawback [84] developed a multi objective incremental learning evolution based fuzzy clustering algorithm for clustering categorical data. The evolution based fuzzy clustering method was combined with random forest classifier for categorical clustering. This algorithm was tested for "synthetic and real world datasets to show the performance of the algorithm".

In many SOM, categorical data cannot be directly processed. It should be converted into a binary value before processing. [32] developed a SOM architecture which processes categorical data without any conversion to binary values.

[85] developed a clustering algorithm by combining "rough set and fuzzy set theories". "An ensemble based framework is designed to find the best clustering results for different categorical data sets".

Other approaches

IIOAB JOURNAL ISSN: 0976-3104

[43] used Bernoulli distribution mixtures for the cluster analysis with binary data, and the results were compared with the Monte-Carlo numerical experiments. [90] proposed an extension of Latent class analysis model for improving the clustering accuracy in each cluster and used Bernoulli distribution mixtures to solve the difficulties of the clustering problem, i.e. to find the number of clusters and to find the correlation matrix for each cluster, etc.
CACTUS proposed by [39], is a summarization based clustering method for categorical dataset with large number of attributes. It required only two scan of the dataset for the formation of clusters and it performs subsace clustering to find the clusters in the subset of attributes. It is a three phase algorithm, first is a summarization phase, second one is a clustering phase and the last phase is a validation phase. The performance of CACTUS was tested for real life and synthetic datasets and it was compared with the existing algorithms.

Squeezer a clustering algorithm proposed by [98] for categorical attributes is suitable for clustering data streams. This algorithm is suitable for solving small dataset only. For handling of large datasets, they proposed an enhanced algorithm called d-squeezer. SCLOPE is also a clustering algorithm for categorical data streams proposed by (Ong et al. 2003).

(He et al. 2005) considered the commonalities between the two different research problems, categorical data clustering and the cluster ensembles. They developed an algorithm based on cross-fertilization between a two problems for clustering categorical data. Whereas, [56] proposed a link based approach for solving the above said two problems.

"[79] proposed an algorithm for clustering categorical data based on the rough set theory called min-min roughness (MMR). The MMR can handle the uncertainty in the clustering process.

[68] proposed a hierarchical clustering algorithm for categorical data based on the rough set model. ATMDP (Total Mean Distribution Precision) method for selecting the partitioning attribute based on probabilistic rough set theory also developed. Based on the TMDP a clustering algorithm called MT MDP (Maximum Total Mean Distribution Precision) was developed. The performance of the MT MDP was compared with the MMR algorithm and claimed that MT MDP algorithm was superior to the MMR algorithm.

[73] compared with some of the existing categorical clustering algorithms using Monte Carlo simulation. The algorithms are like average linkage, ROCK, k-modes, fuzzy k-modes and k-populations were compared.

[67] developed a dissimilarity measure termed as CATCH (an effective Categorical data dissimilarity measure using a distributional Characteristic in High-dimensional space) for clustering categorical data. Zhang and Gu (2014) developed a similarity measure and a affinity propagation (AP) algorithm for clustering mixed data types.

[95] developed a k-modes type clustering algorithm for categorical data which improves the quality of the clusters by using non-dominated sorting genetic algorithm-fuzzy membership chromosome (NSGA-FMC) which combines fuzzy genetic algorithm and multi-objective optimization. Park and Choi (2015) proposed a roughest based approach for clustering categorical dataset named information-theoretic dependency roughness (ITDR).

[26] proposed a method called Maximal Resemblance Data Labeling (MARDL) for clustering concept drifting categorical data. For the concept drifting an algorithm named DCD Detecting concept, drift was also developed. The objective of the algorithm was to find the difference between the distributions of the clusters of the current clustering subset and the last subset. It decides whether the re-clustering was required or not. (Reddy etal.2014) developed a method for data labeling and the concept drift detection based on the entropy model in rough set theory.(Li Y et al. 2014) proposed a three dissimilarity measures based on incremental entropy and an integrated framework consists of a three algorithms for clustering categorical data streams with concept drift.

Many subspace clustering algorithms were proposed for clustering categorical datasets. Subspace clustering is used to find clusters within the datasets in different subspaces (Parson et al. 2004). [3, 1,12, 40, 29] developed a subspace clustering algorithm for categorical data.

(Hatamlou 2013) developed an optimization algorithm named Black hole for data clustering. Black hole algorithm also starts with the initial population solutions for an optimization problem like other population-based methods. In all iterations the best candidate was selected to the black hole. (Hautamäki et al. 2014) proposed a novel clustering algorithm based on alocal search for the objective function. The expected entropy was considered as the objective function for this algorithm. The results were compared with the existing six iterative clustering algorithms and showed that the proposed method produced the best clustering results than the other six methods".

Comparison of various clustering methods

[Table 4] describes the comparison of various methods with respect to the following criteria.

IIOAB JOURNAL ISSN: 0976-3104

- K : The number of clusters known apriori.
    - The value in the table is YES if the cluster number is known at the beginning of the algorithm else, the value is NO
- N: Number of datasets solved
- LD: Largest size of the dataset solved
- S: Whether synthetic datasets generated and tested
    - The value in the table is YES if a dataset is generated and tested else, the value is NO
- C: Compared with the existing methods
    - The value in the table is YES if it is compared with the existing methods otherwise, the value is NO
- Software's or programming languages used for implementing the algorithm in various research articles.

From the literature, it is clear that most of the algorithms required the number of clusters as input, very few algorithms only automatically decides the number of clusters. In the same way, many partitioning based clustering algorithms required a random selection or pre-setting of initial points (mean or modes) of the clusters for clustering. Choosing of these initial points randomly will leads to different cluster results. So, the k-populations algorithm emerged to "automatically initialize the cluster centers and the number of clusters, which leads to the better clustering results [88]".Each clustering algorithm has its own merits and demerits. There is no common clustering algorithm available for handling all kinds of data types. One dimensional clustering by considering all the attributes in the dataset for clustering categorical data is not appropriate for complex datasets so the multi-dimensional clustering was proposed by [29].

Table 4: Comparison of various clustering methods

S. No	Source	K	N	LD	S	C	Impl. Tools
1.	Zhang T et al.(1996)	NA	NA	NA	NO	NO	-
2.	Huang Z (1998)	YES	2	690	YES	YES	-
3.	Ganti V et al. (1999)	YES	2	30919	NO	YES	-
4.	Karypis G, Han ES (1999)	YES	5	10000	NO	YES	-
5.	Guha S et al. (2000)	YES	3	8124	YES	YES	-
6.	Barbará D et al. (2002)	YES	3	1000	YES	YES	-
7.	Ng MK, Wong JC (2002)	YES	4	690	NO	YES	C++
8.	Sun Y et al. (2002)	YES	1	47	NO	YES	-
9.	Zengyou H et al. (2002)	NO	2	8124	YES	YES	Java
10.	Szeto LK et al. (2003)	NO	1	6178	NO	YES	-
11.	Andritsos P et al. (2004)	YES	3	8124	YES	YES	-
12.	Kim DW, et al. (2004)	YES	3	202	NO	YES	-
13.	Ong K et al. (2004)	YES	4	990,002	YES	YES	C
14.	Chang CH, Ding ZK (2005)	YES	5	8124	NO	YES	-
15.	Dutta M et al.((2005)	YES	2	8124	NO	YES	-
16.	He Z, Xu X, Deng S (2005)	YES	4	8124	NO	YES	
17.	Kim DW et al.(2005)	YES	4	202	NO	YES	-
18.	Li T (2005)	YES	6	8280	NO	NO	-
19.	Ahmad A, Dey L (2007)	YES	4	690	NO	YES	-
20.	Cesario E et al. (2007)	YES	13	8124	YES	YES	C++
21.	Parmar D et al. (2007)	YES	3	8124	NO	YES	VB.Net
22.	He Z et al.((2008)	YES	3	8124	NO	YES	-
23.	Andreopoulos B et al.((2009)	NO	5	12960	NO	YES	Python

"K- Number of clusters known Apriori; N- Number of real life dataset solved; LD-Largest size of the dataset; S- Synthetic datasets used;

C- Compared with existing algorithms; Impl. Tools- Implementation tools; NA- Not Available"

Table 4: Comparison of various clustering methods (continued)

S. No	Source	K	N	LD	S	C	Impl. Tools
1.	Cao F et al. (2009)	YES	4	8124	NO	YES	-
2.	Chen HL et al.(2009)	NA	NA	493,857	NO	NA	-
3.	Chen K, Liu L (2009)	NO	1	2,458,284	YES	YES	-
4.	Gan G et al. (2009)	YES	2	435	NO	YES	C++
5.	Mukhopadhyay, A et al.((2009)	YES	4	699	YES	YES	MATLAB
6.	Aranganayagi S, Thangavel K (2010)	NO	4	8124	NO	YES	-
7.	Deng S et al.(2010)	YES	4	8124	NO	YES	Java
8.	Tamhane AC, Qiu D, Ankenman BE (2010)	YES	2	10658	YES	YES	C++
9.	Ahmad A, Dey L (2011)	YES	4	8124	NO	YES	-
10.	Bai L et al. (2011)	YES	4	8124	NO	YES	-
11.	Bai L et al. (2011)	YES	7	2,458,284	YES	YES	-
12.	Cao F, Liang J (2011)	YES	1	8124	NO	YES	-

IIOAB JOURNAL ISSN: 0976-3104

13.	He Z et al. (2011)	YES	5	12690	NO	YES	Java
14.	Rendón E et al.(2011)	YES	NA	NA	YES	YES	-
15.	Bai L et al.(2012)	YES	6	67,557	NO	YES	-
16.	Barcelo-Rico F, Diez JL (2012)	YES	6	30161	NO	YES	-
17.	Cao F et al. (2012)	YES	5	12690	NO	YES	MATLAB
18.	Chen T et al. (2012)	YES	32	20000	YES	YES	Java
19.	Hatamlou A (2012)	YES	6	1473	NO	YES	-
20.	Hsu CC, Lin SH (2012)	YES	2	48842	YES	YES	-
21.	Iam-On N et al. (2012)	YES	9	100000	NO	YES	-

"K- Number of clusters known Apriori; N- Number of real life dataset solved; LD-Largest size of the dataset; S- Synthetic datasets used;
C- Compared with existing algorithms; Impl. Tools- Implementation tools; NA- Not Available"

Table 4: Comparison of various clustering methods (continued)

S. No	Source	K	N	LD	S	C	Impl. Tools
1.	Reddy HV et al. (2014)	YES	NA	NA	YES	NO	-
2.	Saha I, Maulik U (2014)	YES	4	690	YES	YES	-
3.	Zhang K, Gu X (2014)	NO	4	690	NO	YES	C
4.	Bai L, Liang J (2015)	YES	12	8124	NO	YES	-
5.	Bakr AM et al. (2015)	YES	6	20000	NO	YES	-
6.	Baudry JP et al. (2015)	YES	3	440	NO	NA	-
7.	Bouguessa M (2015)	YES	3	8124	YES	YES	-
8.	Del Coso C et al. (2015)	YES	5	48842	NO	YES	-
9.	Dos Santos TRL, Zárate LE (2015)	NA	15	893	NO	YES	-
10.	García-Magariños M, Vilar J (2015)	YES	1	6000	YES	YES	R
11.	Park IK, Choi GS (2015)	YES	1	101	NO	YES	MATLAB
12.	Saha I et al. (2015)	YES	4	435	YES	YES	MATLAB
13.	Seman A et al. (2015)	YES	5	699	NO	YES	-
14.	Tang Y et al. (2015)	YES	2	2400	NO	YES	-
15.	Yang CL et al. (2015)	YES	3	435	NO	YES	MATLAB
16.	Chen LiFei et al. (2016)	YES	4	3190	YES	YES	-

"K- Number of clusters known Apriori; N- Number of real life dataset solved; LD-Largest size of the dataset; S- Synthetic datasets used;
C- Compared with existing algorithms; Impl. Tools- Implementation tools; NA- Not Available"

Frequently used datasets

The real life data set repositories available for clustering are "Frequent Item set Mining Dataset Repository (FIMI), University of California Irvine Machine Learning Repository(UCI)", their URL's are given as follows

"FIMI - (http://fimi.cs.helsinki.fi/testdata.html)
UCI – (http://www.ics.uci.edu/mlearn/MLRepository.html)".

[Table 5] shows a ten frequently used real life datasets with the number of objects and the number of attributes.

Table 5: Frequently used real life datasets

S. No.	Datasets	No. of instances	No. of attributes
1	Soybean	47	35
2	Zoo	101	16
3	Heart Disease	303	13
4	Dermatology	366	33
5	Congressional votes	435	16
6	Credit Approval	690	15
7	Wisconsin Breast Cancer	699	9
8	Car evaluation	1728	4
9	Chess	3196	36
10	Mushroom	8124	22

Validation Measures

[82] compared six internal indexes such as Bayesian information criterion (BIC), Calinski-Harabasz (CH), Davies - Bouldin (DB),Silhouette(SIL), Novel Validity index(NIVA) and DUNN index and four external indexes such as purity, Entropy, F-measure, and Normalized mutual index (NMI) for 13 datasets. The clusters for the comparison were obtained by the k-means and Bisecting-K means algorithms and reported that the internal

measures are more accurate than the external measures. [Table 6] and [Table 7] show the external and the internal validation measures used in the evaluation of the clustering in different studies respectively.

Table 6: External validation measures incorporated in various studies

S.No	External Validation measure	Articles
1.	Clustering Accuracy / Purity	[3],[6],[7],[9],[11],[12],[15],[18],[19],[20],[21], [22], [23], [24],[30],[32],[37],[41],[44], [52],[53], [55], [56],[57],[60],[61],[63], [64],[65],[68],[69],[70],[75],[76],[77], [79],[81],[82],[83],[87],[88]
2.	Adjusted Rand Index	[10],[18],[21], [41],[57],[68],[74],[83],[91],[95]
3.	Number of correctly classified instances	[71]
4.	Micro-right	[71]
5.	Confusion matrix	[49],[67]
6.	Normalized Mutual Information	[29],[56], [57],[68],[82]
7.	Precision or Recall or F-measure or micro precision	[1], [4],[7],[9],[11],[12],[20],[25],[31],[32],[46],[49],[82],[87],[99]
8.	Error rate	[10],[23], [37], [38], [47], [55],[65],[98]
9.	Average clustering error	[2], [33],[50], [51]
10.	Gain ratio	[53]
11.	Category utility	[5], [8], [10],[14], [25]
12.	CPU time	[4], [36],[93], [95]
13.	Jaccard	[18]
14.	Fowlkes	[18]
15.	Entropy	[14],[35],[48],[53],[82]

Table 7: Internal validation measures incorporated in various studies

S. No	Internal Validation measure	Articles
1.	Dunn index	[82],[84]
2.	Silhouette	[35],[41],[82]
3.	Davies-Bouldin index(DB)	[82],[84]
4.	Bayesian information criterion(BIC)	[82],[91]
5.	Novel Validity Index(NIVA)	[82]
6.	Calinski-Harabasz index	[82]
7.	Percentage of correct pair (%CP)	[83], [84]
8.	Minkowski score (MS)	[83],[84]
9.	Compactness	[35], [95]
10.	Gavrilov index (GI)	[41]

Tools for performing clustering

There are few software's available for performing some data mining techniques including clustering. Some of the software's are open source, and few are proprietary version. The details of the commonly used software's for clustering are given in [Table 8].

Table 8: Most widely used open source software's

S. No	Software	Type	Year
1.	CLUTO	Open Source	2002
2.	gCLUTO	Open Source	2003
3.	MALLET	Open Source	2011
4.	Mlpy	Open Source	2012
5.	Orange	Open Source	2009
6.	R-'cluster' Package- CRAN, Rattle	Open Source	2011
7.	TANAGRA	Open Source	2004
8.	wCLUTO	Open Source	2003
9.	Waikato Environment for Knowledge Analysis (Weka)	Open Source	1993
10.	MATLAB- Clustering tool box	Proprietary	1984
11.	Origin	Proprietary	1993
12.	RapidMiner	Proprietary	2001
13.	Statistical Analysis System(SAS)	Proprietary	1971
14.	SPSS	Proprietary	1968

CONCLUSION

This paper provides the overview of the methods used for clustering categorical clustering data like similarity or dissimilarity measures, validation measures available in the literature, and available real life categorical datasets experimented in the different studies. Most of the authors incorporated partition based and hierarchical based methods for clustering categorical data. Partition based clustering is suitable for all types of data, the only drawback of this method is, the number of clusters must be known apriori. This may overcome by choosing the random number of clusters and then increase or decrease the number of clusters to certain level based on the accuracy. Subspace clustering method is appropriate for clustering high-dimensional categorical data. There are very few algorithms only available for model-based and density-based clustering methods and also few papers only incorporated the optimization techniques for categorical data clustering. A small number of heuristics or meta heuristics methods only available for clustering categorical data. The scope for the future research includes the formation of algorithms based on evolutionary algorithms like Genetic Algorithm, Simulated Annealing, and etc. for clustering categorical data.

CONFLICT OF INTEREST
There is no conflict of interest.

ACKNOWLEDGEMENTS
None

FINANCIAL DISCLOSURE
None

REFERENCES

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

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

[3] Al-Razgan M, Domeniconi C, Barba D. [2008] Random subspace ensembles for clustering categorical data. Supervised and Unsupervised Ensemble Methods and their Applications. Studies in Computational Intelligence. 126:31- 48.

[4] Andreopoulos B, An A, Wang X, Labudde D. [2009] Efficient layered density-based clustering of categorical data. Journal of Biomedical Informatics. 42:365–376.

[5] Andritsos P, Tsaparas P, Miller RJ, Sevcik KC. [2004] LIMBO: Scalable clustering of categorical data. Advances in Database Technology-EDBT. 2992:123–146.

[6] Aranganayagi S, Thangavel K. [2010] Incremental Algorithm to Cluster the Categorical Data with Frequency Based Similarity Measure. International Journal of Information and Mathematical Sciences. 6: 21-29.

[7] Bai L, Liang J. [2014] The k-modes type clustering plus between-cluster information for categorical data Neurocomputing. 133:111–121.

[8] Bai L, Liang J. [2015] Cluster validity functions for categorical data: a solution-space perspective. Data Mining and Knowledge Discovery. 29:1560–1597.

[9] Bai L, Liang J, Dang C. [2011] an initialization method to simultaneously find initial cluster centers and the number of clusters for clustering categorical data. Knowledge-Based Systems. 24:785–795.

[10] Bai L, Liang J, Dang C, Cao F. [2011] A novel attribute weighting algorithm for clustering high-dimensional categorical data. Pattern Recognition. 44:2843–2861.

[11] Bai L, Liang J, Dang C, Cao F. [2012] A cluster centers initialization method for clustering categorical data. Expert Systems with Applications. 39:8022–8029.

[12] Bai L, Liang J, Dang C, Cao F. [2013] A novel fuzzy clustering algorithm with between-cluster information for categorical data. Fuzzy Sets and Systems. 215:55–73.

[13] Bakr AM, Ghanem NM, Ismail MA. [2015] efficient incremental density-based algorithm for clustering large datasets. Alexandria Engineering Journal. 54:1147–1154.

[14] Barbará D, Couto J, Li Y. [2002] COOLCAT: An entropy-based algorithm for categorical clustering In: Proceedings of11th ACM international conference on information and knowledge management, McLean, VA, USA. 582-589.

[15] Barcelo-Rico F, Diez JL. [2012] Geometrical codification for clustering mixed categorical and numerical databases. Journal of Intelligent Information Systems. 39:167–185.

[16] Barthelemy J, Brucker F. [2008] Binary clustering. Discrete Applied Mathematics. 156:1237–1250.

[17] Baudry JP, Cardoso M, Celeux G, Amorim MJ, Ferreira AS. [2015] Enhancing the selection of a model-based clustering with external categorical variables. Advances in Data Analysis and Classification. 1:1–20.

[18] Bouguessa M. [2015] Clustering categorical data in projected spaces. Data Mining and Knowledge Discovery. 29:3–38.

[19] Cao F, Liang J. [2011] A data labeling method for clustering categorical data. Expert Systems with Applications. 38:2381–2385.

[20] Cao F, Liang J, Bai L. [2009] A new initialization method for categorical data clustering. Expert Systems with Applications. 36:10223–10228.

[21] Cao F, Liang J, Li D, Zhao X. [2013] A weighting k-modes algorithm for subspace clustering of categorical data. Neurocomputing. 108:23–30.

[22] Cao F, Liang J, Li D, Bai L, Dang C. [2012] A dissimilarity measure for the k-Modes clustering algorithm. Knowledge-Based Systems. 26:120–127.

[23] Cesario E, Manco G, Ortale R. [2007] Top-Down Parameter-Free Clustering of High-Dimensional Categorical Data. IEEE Transactions on Knowledge and Data Engineering. 19:1607–1624.

[24] Chang CH, Ding ZK. [2005] Categorical data visualization and clustering using subjective factors. Data and Knowledge Engineering. 53:243–262.

[25] Chen HL, Chen MS, Lin SC. [2009] Catching the trend: A framework for clustering concept-drifting categorical data. IEEE Transactions on Knowledge and Data Engineering. 21:652–665.

[26] Chen K, Liu L. [2009] “Best K”: critical clustering structures in categorical datasets. Knowledge and Information Systems. 20:1–33.

[27] Chen LiFei. [2015] A probabilistic framework for optimizing projected clusters with categorical attributes. Science China information sciences. 58:1-15.

[28] Chen LiFei, Wang S, Wang K, Zhu J. [2016] Soft subspace clustering of categorical data with probabilistic distance. Pattern Recognition. 51:322–332.

[29] Chen T, Zhang NL, Liu T, Poon KM, Wang Y. [2012] Model-based multidimensional clustering of categorical data. Artificial Intelligence. 176:2246–2269.

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

[31] Çilingirtürk AM, Ergüt Ö. [2014] Hierarchical Clustering with Simple Matching and Joint Entropy Dissimilarity Measure. Journal of Modern Applied Statistical Methods. 13:329-338.

[32] Del Coso C, Fustes D, Dafonte C, Nóvoa FJ, Rodríguez-Pedreira JM, Arcay B. [2015] Mixing numerical and categorical data in a Self-Organizing Map by means of frequency neurons. Applied Soft Computing. 36:246–254.

[33] Deng S, He Z, Xu X. [2010] G-ANMI: A mutual information based genetic clustering algorithm for categorical data. Knowledge-Based Systems. 23:144–149.

[34] Do H J, Kim JY. [2009] Clustering categorical data based on combinations of attribute values. International Journal of Innovative Computing. Information and Control. 5:4393–4405.

[35] Dos Santos TRL, Zárate LE. [2015] Categorical data clustering: What similarity measure to recommend. Expert Systems with Applications. 42:1247–1260.

[36] Dutta M, Mahanta AK, Pujari AK. [2005] QROCK: a quick version of the ROCK algorithm for clustering of categorical data. Pattern Recognition Letters. 26:2364–2373.

[37] Elavarasi SA, Akilandeswari J. [2014] Occurrence based categorical data Clustering using cosine and binary matching Similarity measure. Journal of Theoretical and Applied Information Technology. 68:209-214.

[38] Gan G, Wu J, Yang Z. [2009] A genetic fuzzy k-Modes algorithm for clustering categorical data. Expert Systems with Applications. 36:1615–1620.

[39] Ganti V, Gehrke J, Ramakrishnan R. [1999] CACTUS–Clustering Categorical Data Using Summaries. In: Proceedings of the 5th ACM SIGKDD international conference on knowledge discovery and data mining, San Diego, CA, USA. 73–83.

[40] Gao C, Pedrycz W, Miao D. [2013] Rough subspace-based clustering ensemble for categorical data. Soft Computing. 17:1643–1658.

[41] García-Magariños M, Vilar J. [2015] A framework for dissimilarity-based partitioning clustering of categorical time series. Data Mining and Knowledge Discovery. 29:466–502.

[42] Gebhardt F. [1999] Cluster tests for geographical areas with binary data. Computational Statistics and Data Analysis. 31:39–58.

[43] Govaert G, Nadif M. [1996] Comparison of the mixture and the classification maximum likelihood in cluster analysis with binary data. Computational Statistics & Data Analysis. 23:65–81.

[44] Guha S, Rastogi R, Shim K. [2000] ROCK: A Robust Clustering Algorithm for Categorical Attributes. Information Systems. 25:345–366.

[45] Han J, Kamber M, Pie J. [2011] Data Mining concepts and techniques, 3rd edition, Morgan Kaufmann publishers Inc. San Francisco, CA, USA.

[46] Hatamlou A. [2012] In search of optimal centroids on data clustering using a binary search algorithm. Pattern Recognition Letters. 33:1756–1760.

[47] Hatamlou A. [2013] Black hole: A new heuristic optimization approach for data clustering. Information Sciences. 222:175–184.

[48] Hautamäki V, Pöllänen A, Kinnunen T, Lee KA, Li H, Fränti P. [2014] A Comparison of Categorical Attribute Data Clustering Methods.LNCS. 8621:53–62.

[49] He X, Feng J, Konte B, Mai ST, Plant C. [2014] Relevant Overlapping Subspace Clusters on Categorical Data Categories and Subject Descriptors. In: Proceedings of the 21th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD). 213–222.

[50] He Z, Xu X, Deng S. [2005] A cluster ensemble method for clustering categorical data. Information Fusion. 6:143–151.

[51] He Z, Xu X, Deng S. [2008] k-ANMI: A mutual information based clustering algorithm for categorical data. Information Fusion. 9:223–233.

[52] He Z, Xu X, Deng S. [2011] Attribute value weighting in k-modes clustering. Expert Systems with Applications. 38:15365–15369.

[53] Hsu CC, Lin SH. [2012] Visualized analysis of mixed numeric and categorical data via extended self-organizing map. IEEE Transactions on Neural Networks and Learning Systems. 23:72–86.

[54] Huang W, Pan Y, Wu J. [2012] Goodman-Kruskal measure associated clustering for categorical data. International Journal of Data Mining, Modeling and Management. 4: 334-360.

[55] Huang Z. [1998] Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. Data Mining and Knowledge Discovery. 2:283–304.

[56] Iam-On N, Boongeon T, Garrett S, Price C. [2012], A link-based cluster ensemble approach for categorical data clustering. IEEE Transactions on Knowledge and Data Engineering. 24:413–425.

[57] Ienco D, Pensa RG, Meo R. [2012] From Context to Distance. ACM Transactions on Knowledge Discovery from Data. 6:1–25.

[58] Jain A, Murty M, Flynn F. [1999] Data Clustering: A Review, ACM Computing Surveys. 31:264-323.

[59] Ji J, Pang W, Zheng Y, Wang Z, Ma Z. [2015] An Initialization Method for Clustering Mixed Numeric and Categorical Data Based on the Density and Distance. International Journal of Pattern Recognition and Artificial Intelligence. 29:1550024.

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

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

[62] Karypis G, Han ES. [1999] CHAMELEON : A Hierarchical Clustering Algorithm Using Dynamic Modeling. IEEE Computer. 32:68–75.

[63] Kim DW, Lee KH, Lee D. [2004] Fuzzy clustering of categorical data using fuzzy centroids. Pattern Recognition Letters. 25:1263–1271.

[64] Kim DW, Lee K, Lee D, Lee KH. [2005] A k-populations algorithm for clustering categorical data. Pattern Recognition. 38:1131–1134.

[65] Kuo RJ, Huang YD, Lin CC, Wu YH, Zulvia FE. [2014] Automatic kernel clustering with bee colony optimization algorithm. Information Sciences. 283:107–122.

[66] Lebbah M, Benabdeslem K. [2010] Visualization and clustering of categorical data with probabilistic self-organizing map. Neural computing and applications. 19:393-404.

[67] Lee J, Lee YJ. [2014] An effective dissimilarity measure for clustering of high-dimensional categorical data. Knowledge and Information Systems. 38:743–757.

[68] Li M, Deng S, Wang L, Feng S, Fan J. [2014] Hierarchical clustering algorithm for categorical data using a probabilistic rough set model. Knowledge-Based Systems. 65:60–71.

[69] Li T. [2005] A general model for clustering binary data. In: Proceeding of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining - KDD '05, 188.

[70] Li Y, Li D, Wang S, Zhai Y. [2014] Incremental entropy-based clustering on categorical data streams with concept drift. Knowledge-Based Systems. 59:33–47.

[71] Li-Na W, Qian L, Yuan Z. [2013] A fuzzy centroids clustering algorithm with between-cluster information for categorical data. Information technology journal. 12:5482-5486.

[72] Mampaey M, Vreeken J. [2013] Summarizing Categorical data by clustering attributes. Data Mining and Knowledge Discovery. 26:130–173.

[73] Mingoti SA, Matos R. [2012] Clustering Algorithms for Categorical Data: A Monte Carlo Study. International Journal of Statistics and Applications. 2:24-32.

[74] Mukhopadhyay A, Maulik U, Bandyopadhyay S. [2009] Multi objective Genetic Algorithm-Based Fuzzy Clustering of Categorical Attributes. IEEE Transactions On Evolutionary Computation. 991-1005.

[75] Ng MK, Wong JC. [2002] Clustering categorical data sets using Tabu search techniques. Pattern Recognition. 35:2783–2790.

[76] Ong K, Li W, Ng W, Lim E. [2004] An Algorithm for Clustering Data streams of categorical attributes. Data warehousing and knowledge discovery Lecture notes in computer science. 3181:209–218.

[77] Park IK, Choi GS. [2015] Rough set approach for clustering categorical data using information-theoretic dependency measure. Information Systems. 48:289–295.

[78] Park SSH, Song JJ, Lee JJH, Lee W, Ree S. [2015] How to measure similarity for multiple categorical data sets. Multimedia Tools and Applications. 74:3489–3505.

[79] Parmar D, Wu T, Blackhurst J. [2007] MMR: An algorithm for clustering categorical data using Rough Set Theory. Data & Knowledge Engineering. 63:879–893.

[80] Qin H, Ma X, Herawan T, Zain JM. [2014] MGR: An information theory based hierarchical divisive clustering algorithm for categorical data. Knowledge-Based Systems. 67:401–411.

[81] Reddy HV, Raju SV, Kumar BS, Jayachandra C. [2014] An Approach for Data Labeling and Concept Drift Detection Based on Entropy Model in Rough Sets for Clustering Categorical Data. Journal of Information & Knowledge Management. 13:1450020.

[82] Rendón E, Abundez I, Arizmendi A, Quiroz EM [2011] Internal versus External cluster validation indexes. International Journal of Computers and Communications 5:27–34

[83] Saha A, Das S. [2015] Categorical fuzzy k-modes clustering with automated feature weight learning. Neurocomputing. 166:422–435.

[84] Saha I, Maulik U. [2014] Incremental learning based multi objective fuzzy clustering for categorical data. Information Sciences. 267:35–57.

[85] Saha I, Sarkar JP, Maulik U. [2015] Ensemble based rough fuzzy clustering for categorical data. Knowledge-Based Systems. 77:114–127.

[86] Seman A, Abu Bakar Z, Mohd. Sapawi A, Othman IR. [2013] A medoids-based method for clustering categorical data. Journal of Artificial Intelligence. 6:257-265.

[87] Seman A, Sapawi AM, Salleh MZ. [2015] Performance Evaluations of κ-Approximate Modal Haplotype Type Algorithms for Clustering Categorical Data. Research Journal of Information Technology. 7:112–120.

[88] Sun Y, Zhu Q, Chen Z. [2002] An iterative initial-points refinement algorithm for categorical data clustering. Pattern Recognition Letters. 23:875–884.

[89] Szeto LK, Liew AWC, Yan H, Tang S. [2003] Gene expression data clustering and visualization based on a binary hierarchical clustering framework. Journal of Visual Languages & Computing. 14:341–362.

[90] Tamhane AC, Qiu D, Ankenman BE. [2010] A Parametric Mixture Model for Clustering Multivariate Binary Data. Statistical Analysis and Data Mining. 3-19.

[91] Tang Y, Browne RP, McNicholas PD. [2015] Model based clustering of high-dimensional binary data. Computational Statistics & Data Analysis. 87:84–101.

[92] Xiang Z, Islam MZ. [2014] The Performance of Objective Functions for Clustering Categorical Data. Knowledge Management and Acquisition for Smart Systems and Services Lecture notes in computer science. 16-28.

[93] Xiong T, Wang S, Mayers A, Monga E. [2012] DHCC: Divisive hierarchical clustering of categorical data. Data Mining and Knowledge Discovery. 24:103–135.

[94] Xu R, Wunsch II D. [2005] Survey of clustering algorithms. IEEE Transactions on Neural Networks. 16:645–678.

[95] Yang CL, Kuo RJ, Chien CH, Quyen NTP. [2015] Non-dominated sorting genetic algorithm using fuzzy membership chromosome for categorical data clustering. Applied Soft Computing. 30:113–122.

[96] Yang Z, Sun X, Hardin JW. [2012] Confidence intervals for the difference of marginal probabilities in clustered matched-pair binary data. Pharmaceutical Statistics.11:386–393.

[97] Yang Z, Sun X, Hardin JW. [2012] Testing non-inferiority for clustered matched-pair binary data in diagnostic medicine. Computational Statistics & Data Analysis. 56:1301–1320.

[98] Zengyou H, Xiaofei X, Shengchun D. [2002] Squeezer: An Efficient Algorithm for Clustering Categorical Data. J. Comput. Sci. &: Technol. 17:611-624.

[99] Zhang K, Gu X. [2014] An Affinity Propagation Clustering Algorithm for Mixed Numeric and Categorical Datasets. Mathematical Problems in Engineering. 1–8.

[100] Zhang T, Ramakrishnan R, Livny M. [1996] BIRCH: An Efficient Data Clustering Databases Method for Very Large. ACM SIGMOD International Conference on Management of Data. 1:103–11.