Turkish Journal of Computer and Mathematics Education Vol.12 No.11 (2021), 894-907

Research Article

Differential Evolution based Cluster optimization for Multi valued data sets

P Gopala Krishna[1], D Lalitha Bhaskari[2]

[1]Research scholar, Dept of CS&SE, AU College of Engineering(A), Andhra University,
[2]Professor, Dept of CS&SE, AU College of Engineering(A), Andhra University,
[2]gopalakrishna.aucsse@gmail.com,[2]lalithabhaskari@yahoo.co.in

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 10 May 2021

Abstract: In data analysis, items were mostly described by a set of characteristics called features,in which each feature contains only single value for each object. Even so, in existence, some features may include more than one value, such as a person with different job descriptions, activities, phone numbers, skills and different mailing addresses. Such features may be called as multi-valued features, and are mostly classified as null features while analyzing the data using machine learning and data mining techniques. In this paper, a proximity function is described to find the proximity between two substances with multi-valued features that are put into effect for Clustering. This distance measure approach allows iterative measurements of the similarities around objects as well as their characteristics. For facilitating the most suitable multi-valued factors, we put forward a model targeting at determining each factor's relative prominence for diverse data extracting problems. The proposed model is an evolutionary strategy that uses Differential strategy for evolutions, which is using the degree of member ship as fitness function. The proposed clustering algorithm as multi valued attribute data cluster optimization based on the strategy of Differential evolution (MVA-DE). Therefore this becomes feasible using any mechanisms for cluster analysis to group similar data. The effectiveness of our model is evinced by performance analysis carried through experimental study. The outcomes of the experiments carried on proposed model were compared with other strategic clustering approaches like fuzzy c- means based Clustering of Multivalued Attribute Data (FCM-MVA) and K-Means with Tanimoto based multi-valued data clustering. The findings demonstrate that our test not only improves the performance the traditional measure of similarity but also outperforms other clustering algorithms on the multi-valued clustering framework.

Keywords: Multivalued features, fuzzy c- means Clustering, k-means Clustering, Differential evolution, Tanimoto measure.

1 Overview

The clustering method is the most focal point of many researchers to contribute and conduct their novel research works, particularly on efficient feature selection. The execution procedure of this method differentiates by depending on the sampled sub-sets of features. Thus, the process of selecting appropriate features is significant, as these are engaged in holding essential information of given data. As depicted in [1], to gain the accuracy in operating and executing the certain data extracting algorithms clustering is significant

The significance of clustering is majorly visualized in various data-sets with multiple dimensionalities. Because data mining techniques needs numerous computational efforts in order to handle various features. According to the existing data mining methods, the representation of any dataset is always in a table format and hence, the features maybe the categorical and arithmetical attributes. The conventional methods reflected weak performance parameter in realistic databases, as these sets mostly include attributes, which can predict several values at a time. For instance, this method is involved in the classification of different types of movies including "horror", "romantic", "documentary", and "action". Depending on the specific database domain, the category of attributes helps in conducting mining procedure. Categorical attributes which are capable to estimate multiple values are hardly impacted by minimizing the dimensionalities of various attributes. Several modern works focused on analyzing efficient membership values for multivalued but these values are not always suitable. Because, the optimal values which are analyzed may be in weak connection with the values of dissimilar attributes. Thus, the selection of an appropriate object towards an optimal cluster for such attributes holding several values is observed.

Few scholars also concentrated on utilizing multi-value attributes for executing clustering with other procedures. As depicted in [2], the selection of attributes is explained using diversified attribute's set with various domain ranges. Even though, these research works failed to explain the selection procedure when attributes capture multiple values at a time. Hence, Multi-Relational Data Mining (MRDM) is an open area for many researchers. It encourages authors to develop effective techniques so as to deal with various databases which include multiple tables, as depicted in [3]. The process of MRDM and its related techniques are deeply described and analyzed in [4], [5], [6], [7], [8], [9], [10] and [11] research works.

In order to decrease the redundancy in the database' key table [12], authors proposed the concept of novel table generation. The new research work focused on specific attribute category that generally denotes attributes which holds multiple values in a dataset and it signifies the perspective of MRDM. Although numerous selection methods are being introduced to determine the attributes with multiple-values, only a few works analyzed the significance of selecting those attributes. Further, a research work [13] introduced a solution, in which, k-feasible values related to specific attribute which is a category of multi-valued in k binary attributed are employed. This permits the approach to employ existing feature selection techniques. However, the major thing that limits this method is that

it involves in enlarging dimensionality of original information and is a big threat to this approach. Hence, with increased value of k, the performance of this procedure might be degraded. To overcome this challenge, this paper focused on describing a modern approach for the clustering the data that entails with multi valued attributes. In regard to this, article also depicts a member ship measure depending on the objects of dataset attributes and also used for cluster optimization.

The upcoming sections of this literature include, Section-2 is all about outline of existing literature. Section-3 focused on optimizing the proposed approach by utilizing a member ship. Section-4 comprises research outcomes and determination procedures of proposed approach. Finally, Section-5 concludes the novel research work and also suggest outlook for future research works in this domain.

2 Review of Research Work

Extracting data with multiple values is highly difficult than extracting single-valued data in terms of process overheads, redundancy, and implementation. Performance degradation of mining techniques is observed in multi-valued mining, as it deals with extremely dense data. To overcome with such complication, research works in [14], [15], [16], [17], [18] and [19] are proposed discretization technique. With the decomposition of sequence of in-finite attributes into a cluster of finite neighboring intervals, the discretization technique decreases the process complexity. Moreover, this technique is highly applicable for such mining algorithms, which completely depends on data volume. In addition, the technique also effectively determines the categorical attributes, as explained in [20] and [21]. The tentative results of this technique represent the benefits including fast process execution and enhanced accuracy rate of learning techniques, as explained in [22]. The work in [22] represents various approaches like Supervised and Unsupervised as explained in [23], [24], [25], [26], Static Approaches and Dynamic Approaches as explained in [27], Local or universal as explained in [26], Splitting and Merging, Direct and Incremental, In direct separation methods, the authors need to evaluate the quantity of k-intervals. Depending on those values, infinite attributes are then divided into k intervals at a time. Incremental approaches begin with easy separation procedure and later on continue to upgrade the process, even though few of attributes requires termination criteria to discontinue the discrete procedure.

In [28], constant width and constant frequency discretization approaches are proposed, which entails unsupervised, universal, direct and static models. The below depicted approaches are few of the most significant techniques under Splitting and Merging categories, Few of the methods explained in [28], [29], [30], [31], [32] and [33] are considered as effective splitting approaches. A noteworthy point is that, with the consideration of empirical studies, CACC described in [32] is efficient than others in terms of performance. On merging methods front, these methods employ a testing procedure for analyzing a point at which specific intervals need to be merged. According to [34], researcher introduced most efficient merging approach. As like other traditional approaches, this algorithm also comprised specific limitations including computing complexity and it requires user participation to define several process parameters and to accomplish the merging procedure.

Giannotti et al. [35] suggested a clustering technique for transnational data using k-means algorithm by using the Jaccard similarity measure to cluster the multi-valued attribute data but meets a weak convergence of the method. Fuyuan Cao. [36] suggested a clustering technique for set-valued data called SV-*k*-modes algorithm here the similarity measure for the two objects with multi-valued attributes is defined and a set-valued mode interpretation of cluster centers is suggested. Wenhao Shu. [37] Proposed a Similarity measure on the unlabeled objects. Subsequently, a features extraction method is designed and characterized by mutual information that is incorporated in a declining universe to speed up the screening process of characteristics. Guha *et al.* [38] offered a ROCK algorithm, which is of the type agglomerative hierarchical clustering method that is unscalable to large data. It is furthermore hard to acquire the interpretable cluster agents from hierarchical clustering results. F. Giannotti, C. Gozzi, [39] in this paper it is described a model of splitting and managing transactions, i.e., it is the representation of discrete data with variable size. Authors adapt the appropriate mathematical separation concept shown in the K- Means method to reflect proximity of transactions, and reshape the group centroid concept in a fine way.

Celebi et al. [40] provided an analysis of clustering strategies for solving the numerical configuration issue. The best k-means clustering being implemented based on the analysis of the most common initialization process. Throughout this study, various massive amounts of data have been used to evaluate the clustering quality. However, the K-means grouping method have other inconveniences, The k-means and the fuzzy Cmeans (FCM) cluster methods by Ghosh and Dubey [41] especially in comparison are premised on their effectiveness in selecting the right data analysis method. This clustering algorithm significantly considered the data in the form of the positions around different input data objects. FCM has been an unsupervised grouping method applied and used

in agricultural, astronomical, biological, environmental, medical imaging, classification and clustering areas, in particular.

2.1 Classification of Information through Attributes with Multiple Values

In data mining, classification of information is one of the challenging functions and has its primary focus on approximate the object class depending on related class attributes.For evaluating similarity measure between different objects, a distance factor is employed. Among all, Euclidean distance [42] is an efficient distance parameter used in this kind of classification algorithm. It deals with numerical attributes. For categorical attributes, distance is computed by assuming variance as One (1) for divergent values and Zero (0) for exact values. If a classification involves several class attributes for the representation of attributes which comprises multiple values then a specific distance parameter should be employed. This metric can able to compare various objects set of a class. Hence, this contribution includes various measures to evaluate the distances amid instances sets. In particular, the works in [43], [44] and [3] are employed for analyzing distance between multi-valued attributes. However, as the outcomes of these three distances metrics are same, this manuscript analyzed the solutions using Tanimoto [44]. The Tanimoto distance between two sets including X to Y is referred as DT (X, Y) and is computed through implementing following formula-

$$D = \frac{|X| + |Y| - 2|X \cap Y|}{|X| + |Y| - |X \cap Y|}$$

A noteworthy point is that, as it depends on intersection range of X and Y sets, the distance metric can be used for distinct data sets. On infinite values front, x_i and y_j are similar, if variance $D = \frac{|x_i|-|y_j|}{|x_i|+|y_j|}$is less than pre-defined threshold value.

3 Methods and Materials

3.1. Proximity measure:

The sequence of determination of the most appropriate values in a multi-valued feature context requires the function of member ship abstraction approaches that mine characteristics as per their participation size and not the number of times they occur. A deeper insight into the working of our method is outlined in the following sections. The approach to find similarity between multi-valued objects while accompanying clustering is dependent on multi-valued characteristic. In association to current metrics it consents much use of more than one point of comparison to find similarity for clustering. In this article the similarity between the objects is found as follows:

Computation of similarity of two multi-valued feature values of the X and Y attributes is represented by DMA(X, Y) and determined by consideration of the distances from two sets of elements, that is, to take into account all possible X and Y pairs of attributes. This can be computed by summing an aggregate of all distances in pairs mentioned in the following mathematical model.

$$\mathrm{DMA(X,Y)} = \begin{cases} \dfrac{\sum_{I=1}^{n}\sum_{j=1}^{m} d(x_i,\ y_j)}{|X||Y|}, & If\ X \neq Y \\ 0 & If\ \mathrm{X = Y} \end{cases} \quad \ldots\ldots\ldots\ldots(1)$$

Where$\mathrm{X} = \{x_1, x_2, x_3 \ldots\ldots\ x_n\}$, n ≥ 1 and$\mathrm{Y} = \{y_1, y_2, y_3 \ldots\ldots\ y_m\}$, m ≥ 1 also $d(x_i, y_j)$ is the distance that is described as below between any couple of values generated from X and Y.

$$d(x_i, y_j) = \begin{cases} |x_i - y_j|, & if\ x_i\ and\ y_j\ are\ continuious\ values \\ 0 & if\ x_i\ and\ y_j\ are\ discrete\ and\ x_i = y_j \\ 1 & if\ x_i\ and\ y_j\ are\ discrete\ and\ x_i \neq y_j \end{cases} \quad \ldots\ldots\ldots\ldots\ldots(2)$$

The$SIM(R_i, R_k)$, Proximity between the two fixed unordered data vectors $R_i\ and\ R_k$that are represented by a set of d number of features is foundby using the similarity between their individual dimensions. The dimension similarity can be calculated by usingDMA(X, Y). From the following equation the $SIM(R_i, R_k)$ can be obtained:

$$SIM(R_i, R_k) = \frac{\sum_{j=1}^{d} \mathrm{DMA}(R_i^j,\ R_k^j)}{d} \quad \ldots\ldots\ldots\ldots\ldots(3)$$

3.2. Fuzzy C-Means (FCM) Algorithm:

Let X be the data set with N objects in which each object is characterized by P number of attributes, where $X = \{X_1, X_2, , \dots, X_N\}$ and each X_i is represented by$X_i = \{A_{i1}, A_{i2}, , \dots, A_{iP}\}$, hence the dataset can be represented by a $N \times P$ matrix. Let the data set X is partitioned into C number of fuzzy clusters by a fuzzy clustering algorithm and also each fuzzy partition is represented by a matrix U in which each element u_{ji} represents the degree of member ship of the object X_i in the cluster j whose values lie between 0 and 1. The Fuzzy C-Means (FCM) is depending on the minimization of the objective function given bellow which corresponds to U, a fuzzy partition C of the data set, the set of centroids$V$.

$$J(X; U, V) = \sum_{j=1}^{C} \sum_{i=1}^{N} (\mu_{ji})^m d^2(X_i, V_j), \quad 2 < C < N ------(4)$$

Where V= $\{V_1, V_2, , \dots, V_C\}$, $V_j \epsilon R^p$ is a centroid of the cluster j that is to be determined. The fuzzy ness of the clusters is determined by the fuzzy index which is $m \epsilon (0, \infty)$. $d^2(X_i, V_j)$ is the distance between X_i *and* V_j which is the inner product metric. The trivial solution problem is eliminated by satisfying the following conditions on U.

$\sum_{j=1}^{C} \mu_{ji} = 1$, $\forall$i and $0 < \sum_{i=1}^{N} \mu_{ji} < N$, $\forall$j

Depending on the above construction the Fuzzy clustering can be done through an iterative optimization of the equation (4).

Fuzzy C-Means (FCM) Algorithm:

1. Choose C and ε;
2. Initialize centroids V_j, $j = 1 \dots \dots C$;
3. Find the degree of member ship of each object in all clusters;

$$\mu_{ji} = \frac{1/\left(d^2(X_i,\ V_j)\right)^{\frac{1}{m-1}}}{\sum_{j=1}^{C} 1/\left(d^2(X_i, V_j)\right)^{\frac{1}{m-1}}} \dots \dots \dots \dots \dots \dots (5)$$

4. The new centroids $\widehat{V_j}$ are computed by

$$\widehat{V_j} = \frac{\sum_{i=1}^{N} (\mu_{ji})^m X_i}{\sum_{i=1}^{N} (\mu_{ji})^m} \dots \dots \dots \dots \dots \dots \dots (6)$$

Also the degree of member ship $\mu_{ji}$is updated to $\widehat{\mu_{jl}}$ according to μ_{ji}

5. If$max_{ji}\left|\mu_{ji} - \widehat{\mu_{jl}}\right| < \varepsilon$, stop otherwise, go to step 4, where $\varepsilon \in (0, 1)$ which is a termination condition.

The FCM algorithm allows each object belongs to each cluster depending on the member ship value that is computed by μ_{ji}. Finally, the algorithm assigns each object to a particular cluster according to the maximum member ship of all clusters. To make use of FCM algorithm for multi-valued data the following construction is made

Let $X = \{X_1, X_2, , \dots, X_N\}$be a set of n multi-valued data. Let data X_j ($1 \le j \le N$) be defined by a set of attributes $\{A_1, A_2, A_3, \dots, A_P\}$ in which the attribute A_l is either a single-valued or multi-valued attribute. EachA_l describes a domain of values denoted by $DMN(A_l) = \{a_l^1, a_l^2, \dots \dots a_l^{n_l}\}$, where n_l is the number of distinct values of attribute A_l for $1 \le l \le P$. If A_l is a single valued attribute then each a_l^i ($1 \le i \le n_l$) is considered as a set of single value and If A_l is a multi-valued attribute then each a_l^i ($1 \le i \le n_l$) is considered as a set of multiple values. A domain $DMN(A_l)$ is defined as a finite and unordered. Let X_j be denoted by$\{x_{j,1}, x_{j,2}, , \dots, x_{j,P}\}$, thus Xj can be logically represented as a conjunction of pairs of attribute-values as given bellow

$[A_1 = x_{j,1}] \wedge [A_2 = x_{j,2}] \wedge \dots \dots \dots \wedge [A_P = x_{j,P}]$ Where $x_{j,l} \epsilon DMN(A_l)$ for $1 \le l \le P$.

The objective of the FCM algorithm for multi-valued data (**FCM-MVA)** is to cluster the data set X into C clusters by minimizing the function as given in the equation $J_m(U, C: X)$.

$$J_m(U, C: X) = \sum_{j-1}^{C} \sum_{i=1}^{N} (\mu_{ij})^m {d_{ij}}^2 \dots \dots \dots \dots \dots \dots \dots (7)$$

Subject to $0 \le \mu_{ij} \le 1$; $1 \le j \le C$; $1 \le i \le N$

$$\sum_{j=1}^{C} \mu_{ij} = 1, \quad i = 1 \dots . N$$

$$0 < \sum_{i=1}^{N} \mu_{ij} < N, \quad j = 1 \dots . C$$

Where μ_{ij} is the membership degree of data Xj to the i[th] cluster which is given bellow in the equation (8), and is additionally an element of a $C \times N$pattern matrix $U = [\mu_{ij}]$.

$$\mu_{ij}(t) = \frac{1}{\sum_{z=1}^{C}(\frac{SIM(V_i,X_j)}{SIM(V_z,X_j)})^{\frac{2}{m-1}}} \dots\dots\dots\dots\dots\dots (8)$$

$V = \{V_1, V_2, , \dots, V_C\}$Consists of the centroids of the fuzzy clusters. Centroid V_i is represented as $\{V_{i1}, V_{i2}, , \dots, V_{iP}\}$ the parameter m controls the fuzziness of membership of each datum.To cluster multi-valued data, the fuzzy k-means algorithm extends to cluster multi-valued data based on the fuzzy c-means-type procedure. First, the method for measuring the distance between a cluster centroid and a datum is proposed, along with the method for updating the cluster centroid at each iteration. The distance measure $SIM(V_i, X_j)$between a centroid Vi and a multi-valued data point Xj is defined as described above in similarity measure which is Eq(3).The cluster centroids are updated when the cluster centroid $V_i = \{V_{i1}, V_{i2}, , \dots, V_{iP}\}$ is given, each $V_{il} \epsilon V_i$ for$1 \leq l \leq P$, based on the type of the attribute. If the attribute A_l is numerical then V_{il} is updated as given bellow.

$V_{il}{}^{t} = \frac{\sum_{i=1}^{n}(\mu_{ij}{}^{(t-1)})^{m} x_i}{\sum_{i=1}^{n}(\mu_{ij}{}^{(t-1)})^{m}}$, j=1… k …… (9)

For the categorical attribute A_l the centroid value V_{il} is updated as given bellow.

$$V_{il}{}^{t} = a_l{}^{(s)} \epsilon\, DMN(A_l) \dots\dots\dots\dots\dots\dots (10)$$

$$where \quad \sum_{x_{jl}=a_l^{(s)}} (\mu_{ij})^m \geq \sum_{x_{jl}=a_l^{(t)}} (\mu_{ij})^m \,, 1 \leq t \leq n_l$$

To make use of Fuzzy C-Means Algorithm for multi-valued attributes which is FCM-MVA we replace the equations 5 by 8 for getting member ship of the objects and equation6 by 9 or 11 to get the updated centroids of the clusters.

3.3. Differential Evolution:

The differential evolution (DE) method [44] is one that has been documented to be vigorous to optimization techniques among the other evolutionary procedures relating to the process of optimization. The DE meaning is identical to Genetic algorithm [45] roughly, but In view of the new genetic variants (new population) it varies with GA. Parent and child chromosomes are often evaluated in terms of fitness, if the child's chromosomes seem to be the most fit, then survive and the parents will be disqualified, if the parent's chromosomes are most fit then children's chromosomes do not survive. Only the parent chromosome is replaced by the most fit child chromosome. That means finally either parents or the fittest among all children whichever is more fit is survived. The various fitness mechanisms and crossover methods adopted by DE illustrate incontemporary literature [46][47][48][49] between the different approaches of different evolution strategies. The research that is investigated regarding DE is discussed in the survey [50].

Creation of Initial Clusters:

The application of FCM-MVA clustering as discussed earlier enables a record to fit into one or perhaps more clusters, where the member ship of the record to the corresponding cluster would be greater than the member ship threshold which is usually greater than or equal to 0.3, In this regard one cluster center may be the other cluster's record.

3.4. Optimization of Clusters using DE (MVA-DE)

The initial clusters will be considered as a set of input chromosomes, and performs Differential Evolution on each set of chromosomes that results pair of new chromosomes (new clusters). Among these input and resultant

Algorithm for cluster optimization using DE:

Let the notation $OCS$is a set representing all possible clusters depicted,
Let the notation NCS is a set contains newly formed clusters after each evolution of the DE algorithm,
$While(OCS \cap NCS \neq OCS)$ Begin

$$NCS = OCS$$

For each new cluster $\{C_i / C_i \in NCS\}$ begin
For each new cluster $\{C_j / C_j \in NCS,\ i \neq j\}$
$COR = \{c_{ij} / c_{ij} \in C_i \cap C_j\}$ //Find all common transactions (crossovers) exists in clusters C_i, C_i as set COR
$NCJ = \emptyset$ // an empty set taken to store the new chromosomes generated from crossover process
Consider $NCJ = \{(C_i, C_j)\}$ // moving the parent chromosomes (clusters) to the set NCJ
For each crossover $\{c_{ijk} / c_{ijk} \in COR\}$ where $1 \leq k \leq |COR|$
Begin
$UC_i = \{c_i \in C_i / member\ ship(c_i) < member\ ship(c_{ijk})\}$ //subset of C_i in which the tuples are predecessor to c_{ijk}
$DC_i = \{c_i \in C_i / member\ ship(c_i) \geq member\ ship(c_{ijk})\}$ //subset of C_i in which the tuples are successor to c_{ijk}
$UC_j = \{c_i \in C_j / member\ ship(c_j) < member\ ship(c_{ijk})\}$ // subset of C_j in which the tuples are predecessor to c_{ijk}
$DC_j = \{c_i \in C_j / member\ ship(c_j) \geq member\ ship(c_{ijk})\}$ // subset of C_j in which the tuples are predecessor to c_{ijk}

$$C_{ijk1} = UC_i \cup DC_j$$

$$C_{ijk2} = UC_j \cup DC_i$$

$NCJ = \{(C_{ijk1}, C_{ijk2})\}$ // moving the pair of child chromosomes (clusters) to the set NCJ
End
Find fitness of each entry which is in NCJ as described in sec 3.4.1
Replace the pair of clusters (C_i, C_j) in NCS by the pair of clusters in NCJ which has maximum fitness.
End
End
$if(OCS \cap NCS \neq OCS)$ Begin
$OCS = \emptyset$ //Empty the set OCS
$OCS = \{C_i / C_i \in NCS\}$
$NCS = \emptyset$ //Empty the set NCS
End
End

chromosomes, fittest pair of chromosomes can survive. The following subsection explores the fitness function used in Differential Evolution Process.

3.4.1. Fitness Function

Consider the given each pair of clusters and find the sum of the membership values of all objects in both of the clusters as depicted in section 3.2. In addition, find the pair of clusters with highest sum of the membership values as an optimal pair of clusters among all pairs of clusters given.

Optimization of Clusters Upon completion of the initial cluster formation process, sort the records in descending order of their degree of member ship of each object in each of the lusters

formed, then perform differential evolution to set back the clusters with maximal fitness, which is given in the following algorithm.

Upon completion of the depicted algorithm, the set NCS contains most optimal clusters projected from multivalued dataset. In order to acclaim the clusters with unique entries, allow records to be the part of only one cluster, such that the respective record should have maximal cluster level membership degree for the corresponding cluster.

4. Simulation Study Phase and Efficiency Observations

In this chapter, empirical studies on datasets, evaluation procedures and related solutions of proposed approach are depicted. In regard to assess the significance of the proposed clustering technique MVA-DE, the experiments also carried on K-means clustering that tends to cluster the given data, the distance measure that used in this regard is Tanimoto distance measure. The Tanimoto distance between two sets including X to Y is referred as D (X, Y) and is computed through implementing following formula-

$$D(X,Y) = \frac{|X| + |Y| - 2|X \cap Y|}{|X| + |Y| - |X \cap Y|}$$

The proximity is assessed on the basis of difference in case of continuous values of x_i and y_j

$$D = \frac{|x_i| - |y_j|}{|x_i| + |y_j|}$$

To assess the significance of the proposed clustering technique MVA-DE, the experiments are also carried on FCM-MVA which is described above. The method has been implemented on a 4-GB RAM capacity and i5 processor machine. For the measurement of the results on the resulting clusters, the scripts are described using Python programming language.

4.1 *The Dataset*

This section explores the projection and properties of the real and synthetic datasets used in experimental study. The real dataset that used in experiments is CORA [52], and the synthetic dataset is generated by hybridizing the projection and volume of the CORA dataset.

4.1.1. Real Dataset

Researchers' focuses on CORA [52] database, as it includes 2,708 data records and plays a prominent role in research. Each data record is a scientific contribution from any of seven types including RL machine learning methods, CBR models, Probabilistic approaches, Rule-based Learning approaches, NNs, Genetic techniques and models based on theory. Each record comprises numerous entries to form a data-subset with 1,433 special words that are referred as attributes. The value set of any two attributes which can hold multiple values are called citing and cited manuscripts.

Each document of CORA includes a sub-set of chosen 5,429 special instance identities as a cluster of Multi-values for such attributes usually involve multiple values. Exactness and level of performance of novel approach is determined by utilizing various cluster determination parameters including cluster pureness and cluster HM and also contradictory concepts of both. So as to setup this, the suggested data files are selected based on topic perspectives, as knowledge bases. In addition, clustering of these files into corpuses is observed to assist the optimal determination of clusters according to the selected parameters.

4.1.2. Synthetic Dataset

The dataset generated, by synthesizing the original CORA dataset by adding additional attributes labeled as keywords, and indexing. In addition, around 2000 additional records included to the original dataset. With the influence of the stated modifications to the CORA dataset, the total records become 4708, the simple attributes remain same with count of 1433, however, the multivalued attributes increased from 2 to 4. Metrics pureness, as well as inverted pureness and HM of cluster takes a prominent role in cluster determination procedure. The category frequency in every resulted cluster termed as purity of cluster [54]. Purity parameter can able to remove noise in the clusters, but it is unable to detect the similarities between the records. For instance, in case, each record is considered as single cluster, then purity parameter assigns higher purity value for those clusters. Thus, inverted purity parameter is implemented and essential for analyzing those data clusters as similar categories. This inverted parameter is important in detecting the cluster, which holds highest recall value for each category.

Determination of a cluster involving every input record gives the highest value to inverted purity due to the fact that, this parameter unable to nullify the combination of various records captured from different categories. A noteworthy point is that HM of document clusters also considered in addition to above two parameters. HM parameter is the inverse purity and combination of purity that estimated by comparing every category with the cluster having higher combined precision and recall [55], [56], [57] termed as F-Measure.

4.2 Statistical and Empirical Study of Proposed Work

The proposed solution ensures optimization of Clusters' which are developed from dataset documents and multi-value features because F-Measure of those clusters is extremely high. Level of purity for each detected cluster will have superior accuracy rates. The below Table 1 depicts the statistical data related to the experimental analysis of proposed solution and the Table 2 represents the outcomes of clustering techniques applied on real dataset CORA.

Table 1: The real dataset Statistics

Number of Documents in CORA [53]	2,708
Number of Simple features	1,433
Count of Complex Attributes (Multi-Valued)	2
The number of clusters	7

Table 2: The outcomes of clustering techniques applied on real dataset CORA

	MVA-DE	FCM-MVA Based clustering	K-Means with Tanimoto based multi-valued data clustering
The average of F-measure	0.91	0.89	0.81
Average Cluster purity	0.93	0.91	0.85
Average Clustering Accuracy	0.89	0.85	0.77

The above results are shown in the following figure 1.

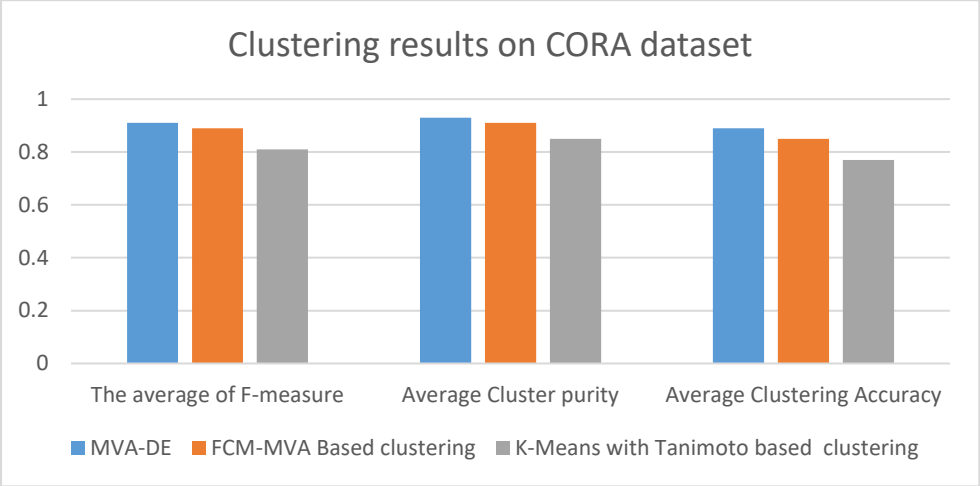

Figure 1: Clustering results on real data set Dataset

In order to further demonstrate the importance of suggested approach, k-means clustering algorithm is implemented on every document along with multi-valued attributes that improve the performance of existing frequency models. The proposed approach also achieves optimal purity and F-Measure parameters. These resulted values of these parameters are effective than the values resulted through earlier methods. The below Figures depicts purity and F-Measure of dissimilar clusters.

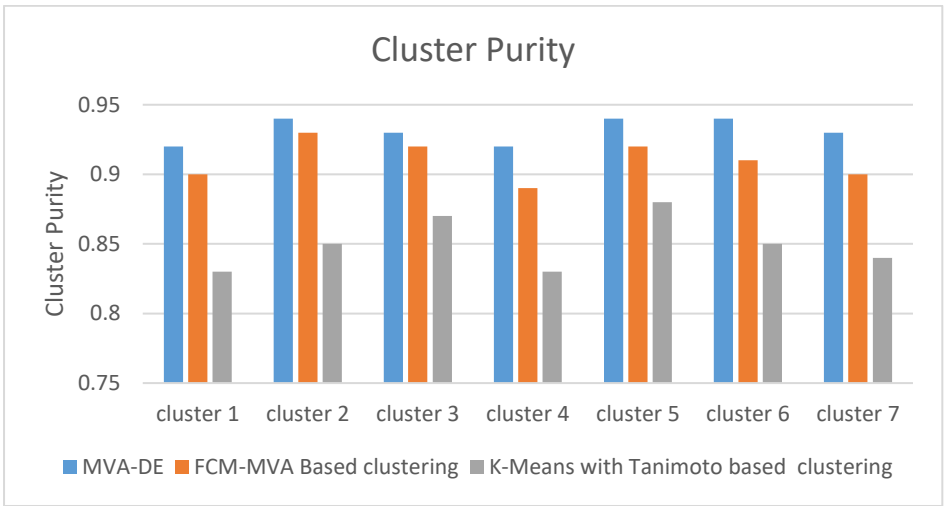

Figure 2: Resulted Purity Value for Dissimilar Clusters

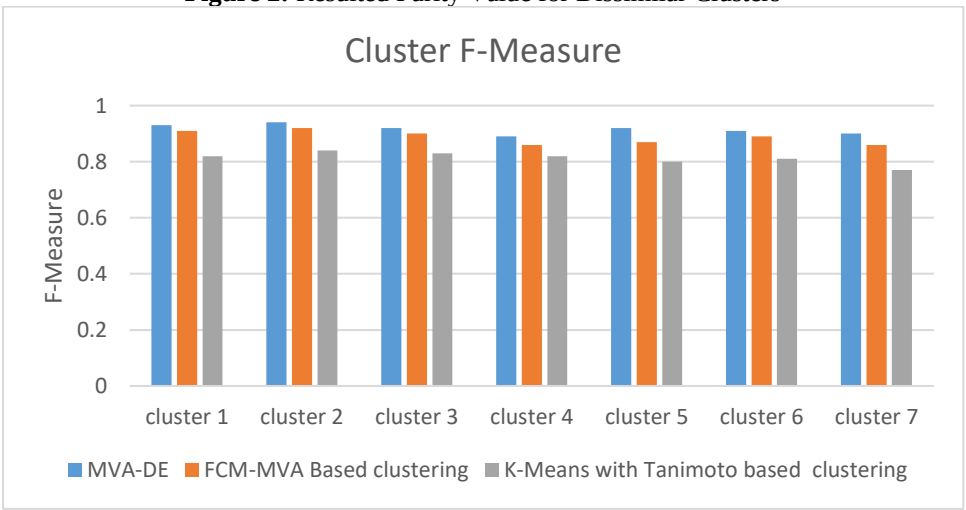

Figure 3: Resulted F-Measure (HM) for Different Clusters

The rate of accuracy visualized for all the approaches is represented in below Figure 4. It represents the reliable proportion value between derived and original true records of an evaluated cluster.

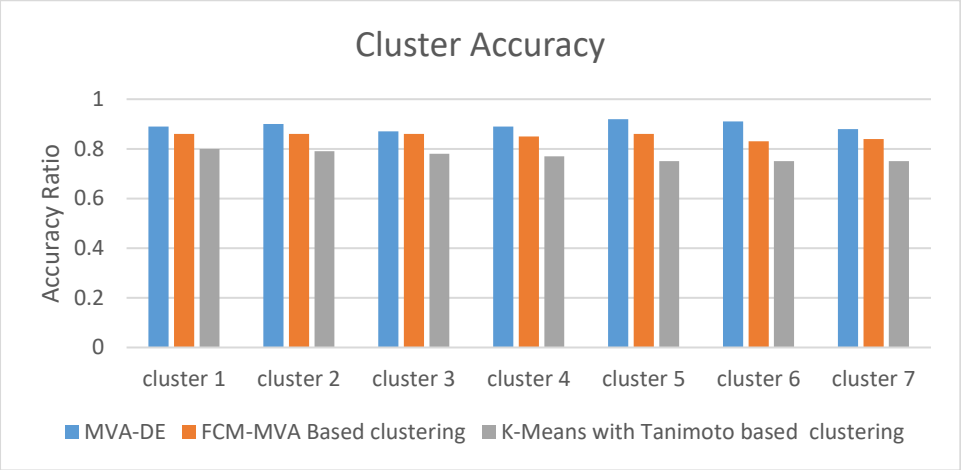

Figure 4: Rate of Accuracy for Dissimilar-Clusters Resulted from all Methods

The similar Assessment is carried on synthetic dataset, the statistics of the dataset are depicted in Table 3, and the performance metric values obtained from proposed and other clustering techniques, those applied on synthetic dataset are depicted in Table 4.

Table 3: The synthetic dataset Statistics

Number of Documents in CORA [53]	4,708
Number of Simple features	1,433
Count of Complex Attributes (Multi-Valued)	41
The number of clusters	7

Table 4: The outcomes of clustering techniques applied on synthetic dataset CORA

	MVA-DE	FCM-MVA Based clustering	K-Means with Tanimoto based multi-valued data clustering
The average of F-measure	0.89	0.87	0.79
Average Cluster purity	0.91	0.89	0.83
Average Clustering Accuracy	0.85	0.82	0.75

The above results are shown in the following figure 5.

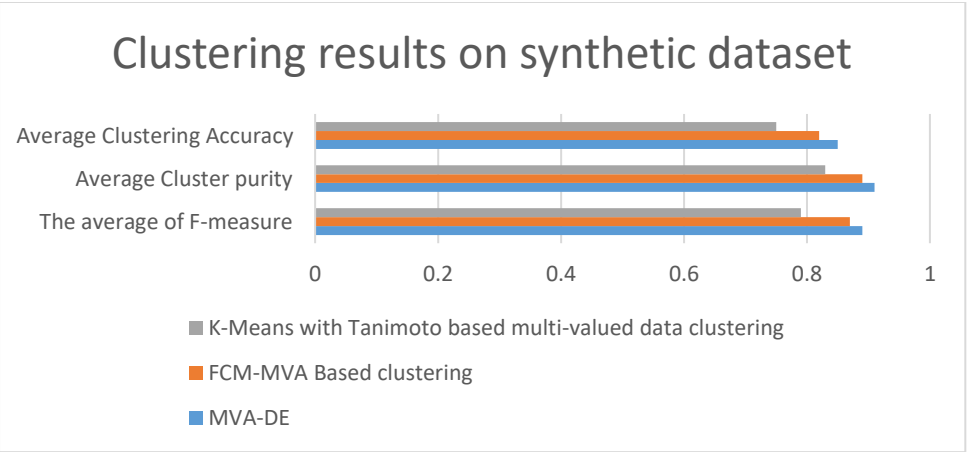

Figure 5: Clustering results on Synthetic Dataset

The results depicted for synthetic data evincing the phenomenal performance advantage of the proposed clustering technique MVA-DE. The resultant clusters purity, accuracy, and cluster harmonic mean observed for DEC-MVA are more than the respective order of k-means clustering with Tanimoto scale-based clustering and FCM-MVA. The cluster level assessment of these three-metrics depicted in Figure 6 (cluster purity), Figure 7 (cluster harmonic means), and Figure 8 (cluster accuracy). It is clearly evincing that all of these cluster level metric values depicted for proposed MVA-DE are stable and outperformed the values depicted for same metrics in regard to other two clustering processes.

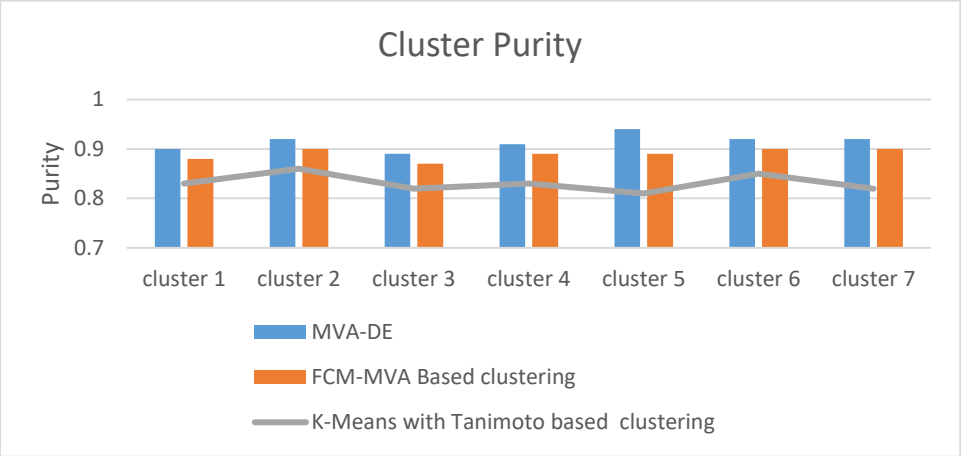

Figure 6: Cluster purity observed for each cluster depicted from proposed and other two clustering models

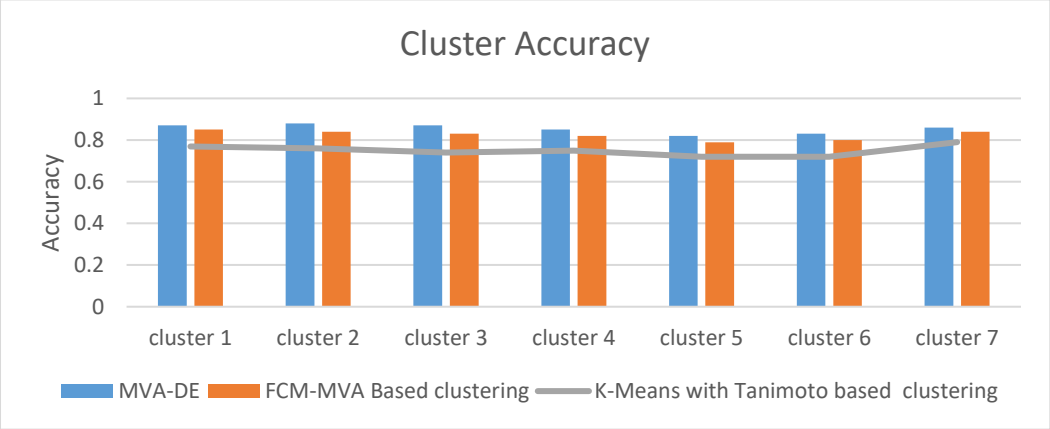

Figure 7: Cluster accuracy observed for each cluster depicted from proposed and other two clustering models

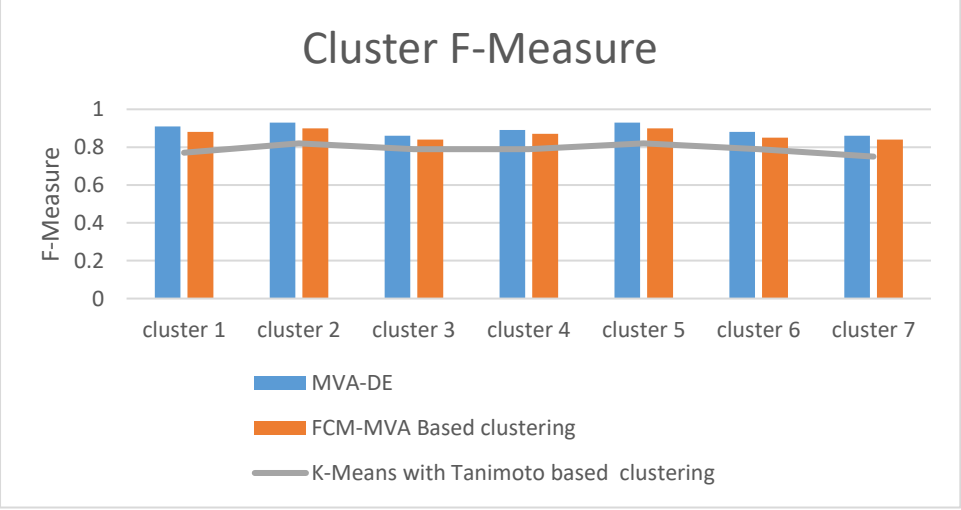

Figure 8: Cluster harmonic mean observed for each cluster depicted from proposed and other two clustering models

5. Conclusion

This contribution proposes a novel approach in order to cluster the data engaged with multivalued attributes. The depicted model is an evolutionary strategy that uses Differential Evolution technique to cluster the data with multivalued attributes. The depicted model is using the degree of membership as fitness measure. In contrast to the selecting methods through available approaches, this paper clusters the data by selecting member ship values based on potentiality of dataset transactions. The proposed solution uses the degree of member ship in mining. This concept allows programmers to form the clusters on the basis of its member ship. The proposed approach also follows the same procedure. Specific values of any tuple are determined through the member ship of that tuple with respect to the cluster in which it belongs. The respective outcomes of this model depict that the novel approach achieves high performance to select efficient values for multi-value features than existing approaches.

To perform empirical study, a real dataset referred as CORA [52], and a synthetic dataset that generated by hybridizing the CORA dataset is employed. Various cluster performance metrics also used such as purity, f-measure, and accuracy. Results observed from empirical study, encouraged the further research work in numerous ways like utilization of member ship in various approaches, ways to innovate additional effective models to select significant values for attributes which comprise multiple values. Finally, the deployment of heuristic scales is also feasible for selecting optimized clusters for these attributes.

References

1. Liu, Huan, and Hiroshi Motoda. "Feature Extraction, Construction and Selection: A Data Mining Perspective." (1998).
2. Deng, Houtao, George Runger, and Eugene Tuv. "Bias of importance measures for multi-valued attributes and solutions." Proceedings of the 21st international conference on Artificial neural networks-Volume Part II. Springer-Verlag, 2011.
3. Džeroski, Sašo. "Multi-relational data mining: an introduction." ACM SIGKDD Explorations Newsletter 5.1 (2003): 1-16.
4. Dehaspe, Luc, and HannuToironen. "Discovery of relational association rules." Relational Data Mining. Springer-Verlag New York, Inc., 2001.
5. Garriga, Gemma, Roni Khardon, and Luc De Raedt. "On mining closed sets in multi-relational data." (2007).
6. Goethals, Bart, Wim Le Page, and Michael Mampaey. "Mining Interesting Sets and Rules in Relational Databases." (2010).
7. Kramer, Stefan, Nada Lavrač, and Peter Flach. "Propositionalization approaches to relational data mining." Relational Data Mining. Springer-Verlag New York, Inc., 2001.
8. Leiva, Héctor Ariel. MRDTL: A multi-relational decision tree learning algorithm. Diss. Iowa State University, 2002.
9. Nijssen, Siegfried, Aida Jimenez, and Tias Guns. "Constraint-Based Pattern Mining in Multi-relational Databases." Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops. IEEE Computer Society, 2011.
10. Koopman, Arne, and Arno Siebes. "Discovering Relational Item Sets Efficiently." Society for Industrial and Applied Mathematics. Proceedings of the SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2008.
11. Spyropoulou, Eirini, Tijl De Bie, and Mario Boley. "Interesting pattern mining in multi-relational data." Data Mining and Knowledge Discovery 3.28 (2014): 808-849.
12. Elmasri, Ramez, and Sham Navathe. "Fundamentals of database systems." (2016).
13. Hall, Mark A., and Geoffrey Holmes. "Benchmarking Attribute Selection Techniques for Discrete Class Data Mining." IEEE Transactions on Knowledge and Data Engineering 15.6 (2003): 1437-1447.
14. Cormen, Thomas H. Introduction to algorithms. MIT press, 2009.
15. Geiser, Jurgen. "Discretization methods with analytical solutions for a convection-reaction equation with higher-order discretizations." International Journal of Computer Mathematics 86.1 (2009): 163-183.
16. Lee, Chien-I., et al. "A Top-Down and Greedy Method for Discretization of Continuous Attributes." Proceedings of the Fourth International Conference on Fuzzy Systems and Knowledge Discovery-Volume 01. IEEE Computer Society, 2007.
17. Mizianty, Marcin J., Lukasz A. Kurgan, and Marek R. Ogiela. "Discretization as the enabling technique for the Naive Bayes and semi-Naive Bayes-based classification." The Knowledge Engineering Review 25.4 (2010): 421.
18. Yang, Ying, Geoffrey I. Webb, and Xindong Wu. "Discretization Methods." Data Mining and Knowledge Discovery Handbook, ISBN 978-0-387-09822-7. Springer Science+ Business Media, LLC, 2010, p. 101 (2010): 101.

19. Zhou, Lu, and B. Yazici. "Discretization Error Analysis and Adaptive Meshing Algorithms for Fluorescence Diffuse Optical Tomography in the Presence of Measurement Noise." IEEE Transactions on Image Processing 20.4 (2011): 1094-1111.
20. Cios, Krzysztof J., and Lukasz A. Kurgan. "CLIP4: Hybrid inductive machine learning algorithm that generates inequality rules." (2004).
21. Clark, Peter, and Tim Niblett. "The CN2 Induction Algorithm." Machine Learning 3.4 (1989): 261-283.
22. Liu, Huan, et al. "Discretization: An Enabling Technique." Data Mining and Knowledge Discovery 6.4 (2002): 393-423.
23. Dougherty, James, Ron Kohavi, and Mehran Sahami. "Supervised and Unsupervised Discretization of Continuous Features." (1995).
24. Ferreira, Artur, and Mário Figueiredo. "Unsupervised Joint Feature Discretization and Selection." Iberian Conference on Pattern Recognition and Image Analysis. Springer, Berlin, Heidelberg, 2011.
25. Jiang, Shengyi, and Wen Yu. "A Local Density Approach for Unsupervised Feature Discretization." Proceedings of the 5th International Conference on Advanced Data Mining and Applications. Springer-Verlag, 2009.
26. Zeng, An, Qi-Gang Gao, and Dan Pan. "A global unsupervised data discretization algorithm based on collective correlation coefficient." Proceedings of the 24th international conference on Industrial engineering and other applications of applied intelligent systems conference on Modern approaches in applied intelligence-Volume Part I. Springer-Verlag, 2011.
27. Wu, QingXiang, et al. "Improvement of Decision Accuracy Using Discretization of Continuous Attributes." Fuzzy Systems and Knowledge Discovery (2006): 674.
28. Chiu, David KY, Andrew KC Wong, and Benny Cheung. "Information Discovery through Hierarchical Maximum Entropy Discretization and Synthesis." (1991): 125-140.
29. Wong, A. K., and D. K. Chiu. "Synthesizing statistical knowledge from incomplete mixed-mode data." IEEE transactions on pattern analysis and machine intelligence 9.6 (1987): 796-805.
30. Kurgan, Lukasz A., and Krzysztof J. Cios. "CAIM discretization algorithm." IEEE transactions on Knowledge and Data Engineering 16.2 (2004): 145-153.
31. Ching, John Y., Andrew K. C. Wong, and Keith C. C. Chan. "Class-Dependent Discretization for Inductive Learning from Continuous and Mixed-Mode Data." IEEE Transactions on Pattern Analysis and Machine Intelligence 17.7 (1995).
32. Tsai, Cheng-Jung, Chien-I. Lee, and Wei-Pang Yang. "A discretization algorithm based on Class-Attribute Contingency Coefficient." Information Sciences 178 (2008): 714-731.
33. Kurgan, Lukasz A., and Krzysztof J. Cios. "Fast Class-Attribute Interdependence Maximization (CAIM) Discretization Algorithm." ICMLA. 2003.
34. Kerber, Randy. "Chimerge: Discretization of numeric attributes." Proceedings of the tenth national conference on Artificial intelligence. Aaai Press, 1992.
35. F. Giannotti, C. Gozzi, and G. Manco, "Clustering transactional data," in *Principles of Data Mining and Knowledge Discovery* (Lecture Notes in Artificial Intelligence), vol. 2431, T. Elomaa*et al.*, Eds. 2002, pp. 175–187.
36. FuyuanCao , Joshua Zhexue Huang, Jiye Liang, Xingwang Zhao , Yinfeng Meng, Kai Feng, and Yuhua Qian, " An Algorithm for Clustering Categorical Data With Set-Valued Features", in IEEE transactions on neural networks and learning systems, vol. 29, no. 10, october 2018.
37. Wenhao Shu and Wenbin Qian, "Mutual Information-based Feature Selection from Set-valued Data", 2014 IEEE 26th International Conference on Tools with Artificial Intelligence.
38. S. Guha, R. Rastogi, and K. Shim, "Rock: A robust clustering algorithm for categorical attributes," in *Proc. 15th Int. Conf. Data Eng.*, Sydney, NSW, Australia, Mar. 1999, pp. 512–521.
39. F. Giannotti, C. Gozzi, and G. Manco, "Clustering transactional data," in *Principles of Data Mining and Knowledge Discovery* (Lecture Notes in Artificial Intelligence), vol. 2431, T. Elomaa*et al.*, Eds. 2002, pp. 175–187.
40. Joshi, A.; Kaur, R.: A review: comparative study of various clustering techniques in data mining. Int. J. Adv. Res. Comput. Sci. Softw. Eng. **3**(3), 67–70 (2013)
41. Ghosh, S.; Dubey, S.K.: Comparative analysis of k-means and fuzzy c-means algorithms. Int. J. Adv. Comput. Sci. Appl. (IJACSA) **4**(4), 35–39 (2013).
42. Deza, Michel Marie, and Elena Deza. "Encyclopedia of distances." Encyclopedia of Distances. Springer, Berlin, Heidelberg, 2009. 1-583.
43. Kalousis, Alexandros, Adam Woznica, and Melanie Hilario. "A unifying framework for relational distance-based learning founded on relational algebra." Technical Report, Computer Science Department, University of Geneva (2006).
44. Duda, R., Hart, P., Stork, D.: Pattern Classification and Scene Analysis. John Willey and Sons, New York (2001).

45. Storn, Rainer, and Kenneth Price. "Differential Evolution–A Simple and Efficient Heuristic for global Optimization over Continuous Spaces." Journal of Global Optimization 11.4 (1997): 341-359.
46. Mitchell, Melanie, Stephanie Forrest, and John H. Holland. "The Royal Road for Genetic Algorithms: Fitness Landscapes and GA Performance." Proceedings of the First European Conference on Artificial Life. 1991.
47. Brest, Janez, and MirjamSepesyMaucec. "Population size reduction for the differential evolution algorithm." Appl Intell 29 (2008): 228-247.
48. Qin, A. K., V. L. Huang, and P. N. Suganthan. "Differential evolution algorithm with strategy adaptation for global numerical optimization." IEEE Trans. Evol. Comput. 2009.
49. Mininno, Ernesto, et al. "Compact differential evolution." IEEE Transactions on Evolutionary Computation 15.1 (2011): 32-54.
50. Islam, S. M., et al. "An adaptive differential evolution algorithm with novel mutation and crossover strategies for global numerical optimization." IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics: a publication of the IEEE Systems, Man, and Cybernetics Society 42.2 (2012): 482-500.
51. Das, S., and P. N. Suganthan. "Differential Evolution: A Survey of the State-of-the-Art." IEEE Transactions on Evolutionary Computation 1.15 (2011): 4-31.
52. https://relational.fit.cvut.cz/dataset/CORA
53. Zhao, Y., and G. Karypis. "Criterion functions for document clustering: Experiments and analysis." (2001).
54. Van Rijsbergen, C. J. "Foundation of Evaluation." Journal of Documentation 30.4 (1974): 365-73.
55. Larsen, Bjornar, and Chinatsu Aone. "Fast and Effective Text Mining Using Linear-time Document Clustering." (1999).
56. Steinbach, M., G. Karypis, and V. Kumar. "A comparison of document clustering techniques." (2000)