AN EFFICIENTFUZZY C-MEANS CLUSTERING ALGORITHM 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,
Email:gopalakrishna.aucsse@gmail.com
[2]Professor, Dept of CS&SE, AU College of Engineering(A), Andhra University, ,
Email: lalithabhaskari@yahoo.co.in

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, it is proposed a proximity function to be described between two substances with multi-valued features that are put into effect for clustering.The suggested distance 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 algorithm is a partition clustering strategy that uses fuzzy c- means clustering for evolutions, which is using the novel member ship function by utilizing the proposed similarity measure. The proposed clustering algorithm as fuzzy c- means based Clustering of Multivalued Attribute Data (FCM-MVA).Therefore this becomes feasible using any mechanisms for cluster analysis to group similar data. 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.

Index Terms*—Multi-valued data set, Multi-valued feature, Fuzzy C-means algorithm, clustering, k-means clustering*

1. Introduction:

The notation to represent the data for analysis of data and mining of data contains a set of n elements $X = \{X_1, X_2, , \dots, X_n\}$with a set of d features$A = \{A_1, A_2, , \dots, A_d\}$. Throughout this model, the data base X is portrayed as a table of n rows and d columns where every row is a specific object and then every column is an attribute whose value with an object would be a single value. This type of database is given as input to the data mining algorithm for analysis and the corresponding results are obtained, but this type of database is very much common and simplified up to the maximum level possible. Often in real life scenarios, attributes in a database might have several values for a tuple, such as a human being with different job roles, activities and skills. All these documentation is common in survey questions, finance, education, telecommunication services, shopping and clinical databases. The most

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

common illustration of data in such applications is shown in the table1.

Table 1: Example of data with Multi-Valued features

Customer ID	Customer Name	Gender	Languages known	Hobbies
1	Jim	M	{Hindi, Urdu}	{Music, Reading}
2	Jack	F	{Telugu, Tamil}	{Sports, Watching TV}
:	:::::::: ::::::	::::: ::::	:::::::::::: ::::::::::	:::::::::::::::: :::::::::::::::
n	Jill	M	{ Hindi, Tamil }	{ Reading, Swimming}

Without loss of sweeping statement, the information in table I can be deliberated as follows. Let **X**={X_1, X_2 …..X_n} be the database where each X_i is a tuple also called as object in the database and each object is represented by a set of d attributes {A_1, A_2, ….. A_d}, where each attribute A_i may be a single valued attribute or a multi-valued attribute. If the attribute$A_j$is single valued attribute then there exist only one value for each tuple X_i($1 \leq i \leq n$)in the database **X,** on the other hand for eachthe multi-valued attributeA_j there exist a non-empty set of values for every tuple X_i ($1 \leq i \leq n$) in the database **X.** To evaluate a set of multi-valued objects like described in Table I, the widely employed approach introduces duplicate features to depict multi-valued attributes. Every unique value of the multi-valued feature is a duplicate feature where its value is 1 if the object seems to have that value in the multi-valued feature; or else, 0 is assigned to the duplicate attribute.Even though duplicate features modify the interpretation of multi-valued features and allow algorithms to perform clustering or classification to study multi-valued data, this approach can lead in the scattering of multi-valued single attribute information is fragmented into several features.As a result the dimensionality of the database is enormously increased and need to preprocess the data by using dimensionality reduction techniques as described in [1], and at the same time the duplicate attributes may have the entries 0 or 1 which causes the existing proximity measures may not produce the accurate results for data analysis.

The rest of this paper is organized as follows: we first introduce some literature used throughout this paper in Section 2. In Section3, the proposed Similarity measureand FCM algorithm is described. In Section 4, we describe our method. Section 5 presents the experimental evidence that demonstrates the advantage of the proposed algorithm. Finally, Section 6 concludes the paper.

2. Related work:

Giannotti et al. [2] 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. [3] 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. [4] 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

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

to speed up the screening process of characteristics.Guha*et al.* [5] 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, [6] 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.[7] 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[8] 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. The research examines the efficiency of the clustering techniques of the FCM in comparison with that of the k-means methodology of clustering. In the study of the K-means and FCM clustering strategies which discussed in Velmurugan[9] for telecommunications connection-related data. The quality of these methodologies has been assessed based mostly on network connection area. This article has said that, contrast to the k-means clustering, the FCM methodology was much more precise and easier to implement.Wang et al [10] have presented the information of attribute proximity for objects (CASO) to do clustering. The interconnected and intra-connected features for the improved accuracy and the learning of complexity are investigated in this research. Entirely focused on attribute types it is classified as two types they are categorical and continuous. This paper concluded also that categorical grouping strategies based on the results were best appropriate for large-size data. In order to group, control association processing and several other data analysis, Mukhopathy et al. [11] presented various evolutionary multi - objective methodologies. Whenever the number of features is high, the key limitation of the binary coding scheme could not be clustered. In addition, this paper explored two separate algorithms for multi - objective functional rule mining which includes MODENAR and MODE, along with this three specific kinds of data are also studied, such as descriptive, numeric and fuzzy methods of data collection. The experiment showed that categorical strategies for data clustering effectively group huge size data with even a wide range of features. Kim et al. [12] suggested that the GK method cluster accuracy index be focused on the relative standard shared value with all feasible fuzzified-cluster couples. Zhang et al. [13] adopted the Pearson correlation In order to calculate the distance and suggested an appropriate function, In order to affect the labeling of a pixel in the close surroundings, Ahmed et al [14] changed its optimal solution as that of the regular FCM algorithm. The updated proposed method improved the efficiency on

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

noisy pictures from traditional FCM techniques. Nonetheless, the way the adjacent knowledge is integrated restricts its access to single functional inputs.

In order to identify the cluster formation number, Zhang et al. [15] proposed a new WGLI utilizing global best affiliation as general properties and the bipartite system extensibility as smaller local domain. Charles Bouveyron and Brunet-Saumard[16] designed a method of established model-based data segmentation approaches. Grouping on models was a common tool known by its probability - based frameworks and versatility. The Hamming distance-based discrete PSO algorithm for classification and recognition in gene sequences was proposed by Haider Banka and Suresh Dara [17]. The test results suggest that the HDBPSO offered the improvement in the proximity calculation while using the hamming distance. In order to maximize the effectiveness of the Fuzzy classifiers, a new feature selection process has been proposed by LyamineHedjazi et al [18] that includes all blended types and higher dimensional relevant data on membership limits. The findings show that the approach contributes to a great improvement in the efficiency of classification of both fuzzy classification and other state-of-the-art classifiers.

2.1.Data Classification grounded on multi-valued attributes:

The classification of data continues to be a critical step in which instances of classes are based on the relevant features that are predicted. K-NN is really a famous classification framework from an MRDM interpretation [19], and is one of the most common lazy models [20]. The model gained its popularity in recent years 1990s, even though the model was introduced in the 1950s [21].Each individual unidentified entity is defined by comparison with current objects throughout the dataset as the basic definition. It allows k value to be associated only with highest correlation. K- The system is fed with the interest. The unknown instance is identified only with largest value on the strength of its label across all objects.Proximity metrics are used to measure the similarities between different objects. Euclidean distance [22] has become one of the popular distance measures commonly applied to certain classification methods which refers to the numerical values. Mostly in case of nominal variables distance is calculated with the assignment of 0 in different kinds and a 1 for completely identical values.

In cases where single valued or multi-valued attributes were included, a special proximity metric concept must be introduced, which is reliable of comparing various sets. Investigators therefore propose different criteria that describes the proximity between pair of sets. In this proposed study it is analyzed that the studies [23], [24], and [25] represents the proximity metric between various attributes and compiled for computation of such analyses. Distance estimation of all study results concluded in identical results and thus findings of the Tanimoto test were reported in this document.

The proximity of a set of values to another set of values using Tanimoto measure is represented as given bellow

$$D = \frac{|X|+|Y|-2|X\cap Y|}{|X|+|Y|-|X\cap Y|} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

In the above measure the intersection between A and B is applied on the distance measure, it can apply to sets with different

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

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

$$D = \frac{|x_i| - |y_j|}{|x_i| + |y_j|} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

A database with multi-valued attributes includes data of different values. Two concepts namely MMC and MMDT were described for multi-valued databases in [26] and [27]the two methodologies are developed by the decision tree approach. The modified edition of MMC is MMDT of these the MMC separates features, whereas the MMDT method additionally improves certain features, to ensure the highest efficiency of classification details. In the general context these methods could not extract the appropriate optimum features from the multi-valued data base.The study [28] describes a new method to choose the best set of values for multi-valued features, which makes it easier to quantify their significance for extraction method. This model suggested to select values based on related transaction weight, in contrast to the general trend of choosing values for multi-valued features depending on the frequency. The developed concept is generated by the utility analysis techniques, in which the values are chosen according to their significance instead of its occurrence.

In the same context, [29] a multifunctional attribute relevance test called RMULT has been developed to estimate the significance for classification including its multi-valued feature. The aim of this metric is to assess the multi-valued classification feature scope. Even so, multiple values are combined with the multi-valued features, so various values of these characteristics correspond to different groupings.A model by LNC.Prakash K [30], a DE-based Multi-valued Attribute Data (DEC-MVA) clustering algorithm was developed to evaluate the relative importance of each factor in relation to various data extraction issues to promote the most appropriate multi-valued factors. This framework developed also an evolutionary method that utilizes a differential evolution framework that incorporates the transaction utility as optimization process.In this framework, the insight of this article represents a new distance metric that matches a multi-valued characteristics of various classes, this measure is suitable for both clustering and classification methods of databaseanalysis.

3. Similarity measure:

Attempting to follow what has been discussed in the preceding sections on distance metrics, essential aspects should be taken when choosing a suitable distance metric for the job of multi-valued clustering. These aspects have included type of analysis as well as the purpose of analysis which, as a result, determine the type of distance metric that will be used. Whenever the purpose is to determine similarity for multi-valued attributes, i.e. the presence of perfect similar patterns would not only be essential, but it is also important to consider the partial similarity as well as the unmatched values of the multi-valued attribute values.The strategy to find similarity between multi-valued objects while conducting clustering is depending on multi-valued characteristic. In comparison to current metrics it allows much use of more than one point of comparison to find similarity for clustering. In this article the similarities of objects is determined as follows:

proximity calculation between two values of multi-valued attribute values X and Y, denoted by DMA(X, Y) and defined by considering the distance between the elements of the two sets (or multi-valued attribute values) i.e. to consider all possible pairs created by X and Y. It can be calculated by taking the average of all

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

distances in pairs which will be given in the following mathematical formulation.

$$\mathrm{DMA}(X,Y) = \begin{cases} \dfrac{\sum_{I=1}^{n}\sum_{j=1}^{m} d(x_i,\ y_j)}{|X||Y|}, & If\ X \neq Y \\ 0 & If\ X = Y \end{cases} \quad \ldots\ldots\ldots\ldots (3)$$

Where$X = \{x_1, x_2, x_3 \ldots\ldots x_n\}$, n ≥ 1 and$Y = \{y_1, y_2, y_3 \ldots\ldots y_m\}$, m ≥ 1 also $d(x_i, y_j)$ is the distance between each pair of values formed from X and Y which is defined as given bellow.

$$d(x_i, y_j) = \begin{cases} |x_i - y_j|, & if\ x_i\ and\ y_j\ are\ continuious\ values \\ 0 & if\ x_i\ and\ y_j\ are\ discrete\ and\ x_i = y_j \\ 1 & if\ x_i\ and\ y_j\ are\ discrete\ and\ x_i \neq y_j \end{cases} \quad \ldots\ldots\ldots\ldots\ldots (4)$$

The Proximity $SIM(R_i, R_k)$ between the two fixed non-ordered data vectors $R_i\ and\ R_k$ which are described by a set of d number of attributes is defined in terms of the similarity between their individual dimensions. The dimension similarity can be measured using$\mathrm{DMA}(X,Y)$. The value of $SIM(R_i, R_k)$ can be obtained from the following equation:

$$SIM(R_i, R_k) = \frac{\sum_{j-1}^{d} \mathrm{DMA}(R_i^j,\ R_k^j)}{d} \quad \ldots\ldots\ldots\ldots\ldots\ldots (5)$$

3.1.Fuzzy C-Means Clustering Algorithm (FCM):

In this section, we briefly define the Fuzzy C-means algorithm.Find a collection of unidentified objects$X = \{X_1, X_2, , \ldots, X_n\}$, Here n is the total number of variations and the dis of the pattern vectors (features).The FCM algorithm is based on reducing the value of an objective function.The objective function tests the partitioning efficiency that separates a dataset into C clusters.The FCM algorithm tests the Partitioning efficiency by measuring the distance from the pattern $X_{i,}$tothe current candidate cluster center $C_j$with the distancefrom pattern X_i**,** to other candidate cluster centers. The objective function is an optimization function, which determines the weighted total of squared errors within the category, which is minimised as follows.

$$J_m(U, C) = \sum_{j-1}^{C}\sum_{i=1}^{n} (\mu_{ij})^m {d_{ij}}^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots (6)$$

Where

n: the number of patterns in ***X***

C: the number of clusters

U: the membership function matrix; the elements of ***U*** are (μ_{ij})

μ_{ij} **:** the value of the membership function of the ith pattern belonging to thejthcluster

d_{ij}**:** the distance from X_i**,** to C_j**,** where$C_j$denotes the cluster center of the jth cluster

m**:** the exponent on $\mu_{ij,}$ to control fuzziness or amount of clusters overlap

The FCM algorithm focuses on minimizing $\boldsymbol{J_m}$subject to the following constraints on ***U:***

$\mu_{ij} \epsilon$ [0, 1], i= 1… n and j= 1…..C

$$\sum_{j=1}^{C} \mu_{ij} = 1,\ \ i = 1 \ldots\ldots n$$

$$0 < \sum_{i=1}^{n} \mu_{ij} < n,\ \ j = 1 \ldots\ldots C$$

Function $J_m(U, C)$A restricted optimization issue is defined, which can be translated by using the Lagrange multiplier technique to an uncontrolled optimization problem.

$$\mu_{ij} = \frac{1}{\sum_{i=1}^{C} (\frac{d_{ij}}{d_{il}})^{\frac{2}{m-1}}}\ , \text{ i=}$$

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

1… n,andj, l=1….C …….. (7)

Where $\mu_{ij} = \begin{cases} 1 & if \quad d_{ij} = 0 \\ 0 & if \quad l \neq j \end{cases}$

$$C_j{}^t = \frac{\sum_{i=1}^{n} (\mu_{ij}{}^{(t-1)})^m x_i}{\sum_{i=1}^{n} (\mu_{ij}{}^{(t-1)})^m},$$

j=1… C …….. (8)

A range of initial cluster centres begins with the FCM algorithm (or ***arbitrary membership values).*** Then, iterates the two functions upgradingμ_{ij} and$C_j{}^t$ at the i[th] iteration until the cluster centers are stable or the objective function in $J_m(U, C)$ converges to a local minimum.
The complete algorithm consists of the following steps:

Step 1: Given a fixed number *C*, initialize the cluster center matrix C^0 by using **a** random generator from the original dataset. Record the cluster centers, set *t=O, m* = 2, and decide ε, where εis a small positive constant.

Step 2: Initialize the membership matrix U^0 by using functions of μ_{ij}.

Step 3: Increase ***t*** by one. Compute the new cluster center matrix (candidate) C[t]by using$C_j{}^t$.

Step 4: Compute the new membership matrix $\boldsymbol{U}^t$by using functions of μ_{ij}.

Step 5: If $\|U^t - U^{t-1}\| <$ εthen stop, otherwise go to **step3.**

4. Fuzzy C-Means Clustering Algorithm for Multi-Valued Data (FCM-MVA):

Let $X = \{X_1, X_2, , \dots, X_n\}$be a set of n multi-valued data. Let data X_j $(1 \leq j \leq n)$ be defined by a set of attributes $\{A_1, A_2, A_3, \dots, A_d\}$ 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 distinctvalues of attribute A_l for $1 \leq l \leq$ d. If A_l is a single valued attribute then each a_l^i $(1 \leq i \leq n_l)$ is considered as a set of single value and If A_l is a multi-valued attribute then each a_l^i $(1 \leq i \leq n_l)$ is considered as a set of multiple values.A domain $DMN(A_l)$ is defined as a finite and unordered. Let X_j be denotedby$\{x_{j,1}, x_{j,2}, , \dots, x_{j,d}\}$, thus Xj can be logicallyrepresented as a conjunction of pairs of attribute-valuesas given bellow $[A_1 = x_{j,1}] \wedge [A_2 = x_{j,2}] \wedge \dots \dots \dots \wedge [A_d = x_{j,d}]$Where$x_{j,l} \epsilon DMN(A_l)$ for $1 \leq l \leq$ d.

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

$$J_m(U, C: X) = \sum_{j-1}^{k} \sum_{i=1}^{n} (\mu_{ij})^m d_{ij}{}^2 \quad \dots \dots \dots \dots \dots \dots \dots \dots . (9)$$

Subject to$0 \leq \mu_{ij} \leq 1$; $1 \leq j \leq k$; $1 \leq i \leq n$

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

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

Where μ_{ij} is the membership degree of data Xj to the i[th] cluster, and is additionally an element of a $k \times n$pattern matrix $U = [\mu_{ij}]$.$V = \{V_1, V_2, , \dots, V_k\}$Consists of the centroids of the fuzzy clusters. Centroid V_i is represented as$\{V_{i1}, V_{i2}, , \dots, V_{id}\}$the

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

parameter m controls the fuzziness of membership of each datum.

The fuzzy k-means algorithm expands to cluster multi-value data based on the Fuzzy c-means-type technique to cluster multi-value data.Next, the approach is introduced to calculate the distance between a cluster centroid and a datum, along with the process of 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(5).

The cluster centroids are updated when the cluster centroid $V_i = \{V_{i1}, V_{i2}, , \dots, V_{id}\}$ is given, each $V_{il} \epsilon V_i$for$1 \le l \le d$, based on the type of the attribute. If the attribute A_l is numerical thenV_{il} is updatedas given bellow.

$$V_{il}{}^{t} = \frac{\sum_{i=1}^{n} (\mu_{ij}{}^{(t-1)})^{m} x_i}{\sum_{i=1}^{n} (\mu_{ij}{}^{(t-1)})^{m}} \quad , \qquad \text{j=1\dots k} \qquad \dots\dots\dots\dots\dots\dots (10)$$

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

$$V_{il}{}^{t} = a_l{}^{(s)} \epsilon\, DMN(A_l) \quad \dots\dots\dots\dots\dots\dots (11)$$

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

4.1.The proposed clustering algorithm for multi-valued data:

To minimize the objective function $J_m(U, C: X)$that is Eq (9) with suggested centroids which are defined in Eq (10) and Eq (11), The proposed algorithm uses the prototype form Fuzzy c-means for multi-valued cluster data.

Step 1: Choose initial centroids given the number of clusters, k, and a selected value of m $V(0)$, $(t = 0)$. Each $V_{il} \epsilon V$ is assigned random membership values for U^t.

Step 2: Compute the i[th] fuzzy cluster for $i = 1 \dots k$. For each X_j :

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

Step 3: Update the fuzzy cluster centroid

$V(t) = \{V_{i1}, \dots, V_{il}, \dots, V_{id}\}$for i = 1, 2… k.For each$V_{il} \epsilon A_l$.

$$V_{il}{}^{t} = \begin{cases} \frac{\sum_{i=1}^{n} \left(\mu_{ij}{}^{(t-1)}\right)^{m} x_i}{\sum_{i=1}^{n} \left(\mu_{ij}{}^{(t-1)}\right)^{m}} & \text{If } A_l \text{ is continuous valued attribute} \\ \begin{cases} a_l{}^{(s)} \epsilon\, DMN(A_l) \text{If } A_l \text{ is categorical valued attribute} \\ where \sum_{x_{jl}=a_l^{(s)}} (\mu_{ij})^{m} \ge \sum_{x_{jl}=a_l^{(t)}} (\mu_{ij})^{m} \ , 1 \le t \le n_l \end{cases} \end{cases}$$

Step 4: Step 4. If there are no improvement in J_m, then stop; otherwi

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

se, set $t \leftarrow t + 1$ also
go to Step 2.

5. Experimental Study and Performance Analysis

In this section, 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 FCM-MVA, 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 A to B is referred as D (A, B) and is computed through implementing following formula-

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

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

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

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.

5.1.The Dataset

This section explores the projection and properties of the real dataset used in experimental study. The real dataset that used in experiments is CORA [31], and the synthetic dataset is generated by hybridizing the projection and volume of the CORA dataset.

Real Dataset

Researchers' focuses on CORA [31] 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, asknowledge bases. In addition, clustering of these files into corpuses is observed to assist the optimal determination of clusters according to the selected parameters.

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 effect of the improvements to the CORA dataset that have been reported,The overall records are 4708, with the number of 1433 the basic attributes remain the same, but the multi-value

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

attributes have risen from 2 to 4.

Proposed Solution Evaluation Parameters and strategies

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. 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. Inverted purity parameters are therefore introduced and are essential to analyse these 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 records 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 purity mixture determined by comparing each segment with the higher cumulative accuracy and recall cluster[32],[33],[34] referred to as F-Measure.

5.2.Statistical and Empirical Study of Proposed Work

The proposed solution guarantees that clusters that are built from dataset documentation and multi-value features are configured because F-Measure is incredibly large for such clusters. The purity standard would have superior precision rates for each observed cluster.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 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. Table 2 below displays the statistical evidence relevant to the experimental study of the solution suggested.

Table 2: The statistics of the input data and results obtained.

	FCM-MVA Based clustering	K-Means with Tanimoto based multi-valued data clustering
Total number of Records in CORA	2708	2708
The maximum number of simple attributes	1433	1433
The maximum size of the multivalued set	5429	5429
The number of multi-valued attributes	2	2
The number of classes(clusters)	7	7
The average of F-measure	0.89	0.81
Average Cluster purity	0.91	0.85
Average Clustering Accuracy	0.85	0.77

The above results are shown in the following figure 1.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

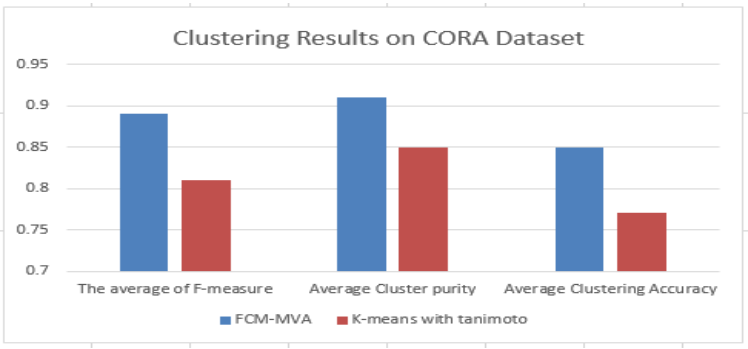

Figure 1: Resulted average Values of Clustering

The below Figures depicts purity and F-Measureof dissimilar clusters.

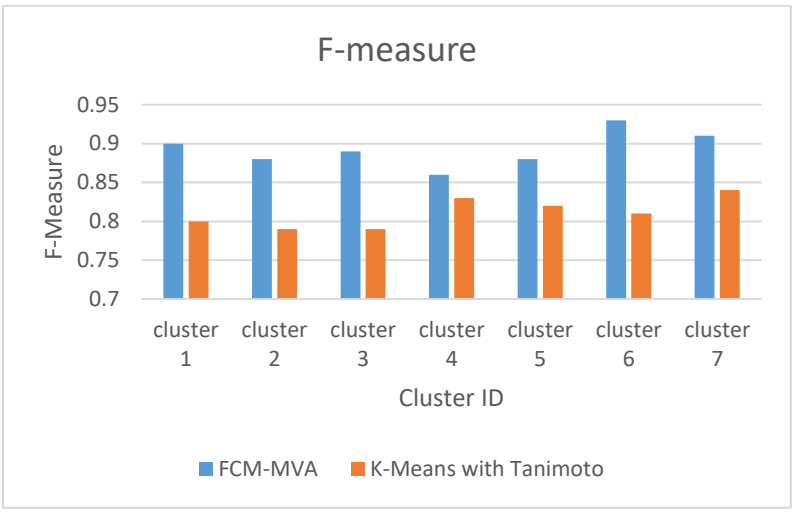

Figure 2: Resulted F-Measure Value for Dissimilar Clusters

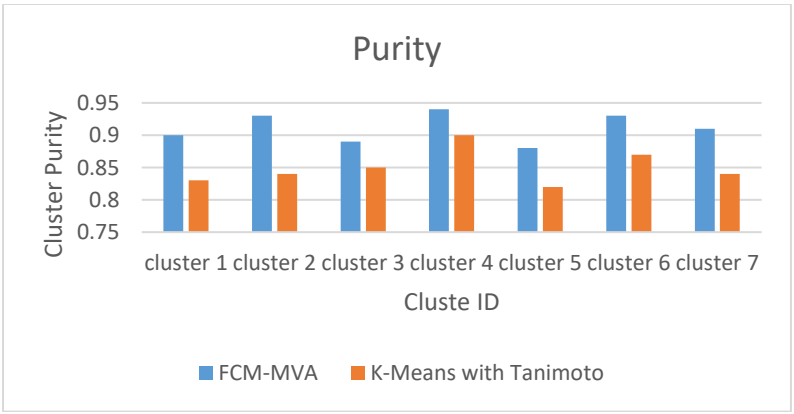

Figure 3: Resulted Purity Value for Dissimilar Clusters

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

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

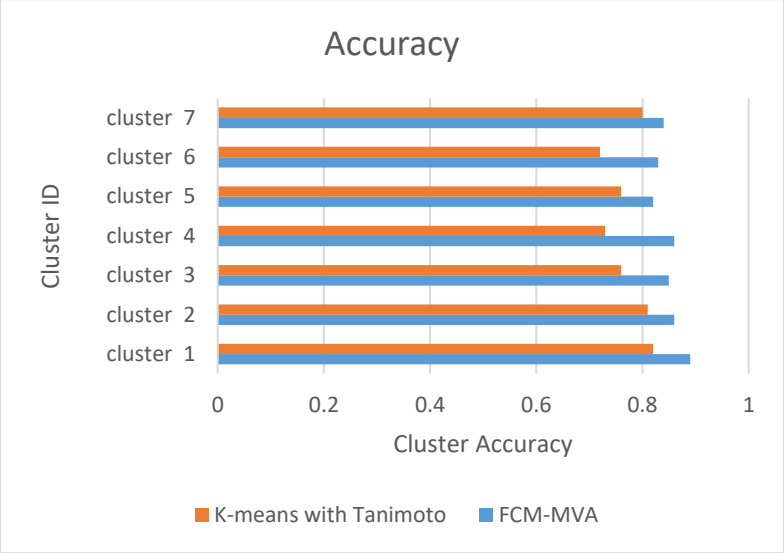

Figure 4: Resulted Accuracy Value for Dissimilar Clusters

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 3.

Table 3: The statistics of the input data and results obtained.

	FCM-MVA Based clustering	K-Means with Tanimoto based multi-valued data clustering
Total number of Records in synthetic dataset CORA	4,708	4,708
The maximum number of simple attributes	1433	1433
The number of multi-valued attributes	4	4
The number of classes (clusters)	7	7
The average of F-measure	0.87	0.79
Average Cluster purity	0.89	0.83
Average Clustering Accuracy	0.82	0.75

The results depicted for synthetic data evincing the phenomenal performance advantage of the proposed clustering technique FCM-MVA. The resultant clusters purity, accuracy, and cluster harmonic mean observed for FCM-MVA are more than the

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

respective order of k-means clustering with Tanimoto based multivalued set optimization.

6. Conclusions

The research work reported in this paper is a first step towards FCM-MVA. Clustering based on non-ordered multi-valued attributes is a key requirement in many data mining applications. Clustering in multi-valued has some extra challenges which are not encountered in mono-valued data. In this paper, we have proposed a similarity measure based on both single valued and multi valued attributes, a clustering technique called FCM-MVA for multi-valued non-ordered discrete and continuous data items. We validated our proposed technique experimentally on synthetic dataset of diverse characteristics as well as with a real dataset. That the experimental results demonstrate the effectiveness of FCM-MVA as a clustering method on a real dataset referred as CORA [31] (Barragao, 2018),consisting of both multiple values and single value attributes are utilized and a synthetic datasetthat generated by hybridizing the CORA dataset is employed.

The experimental analysis also showed the importance of the proposed distance measurement measure for clustering the data under the unsupervised method of learning. The performance review was carried out by evaluating the outcomes of the proposed model and the other contemporary model called "Tanimoto"(Tasca, 2013).Various cluster performancemetricsalso used such as purity, f-measure, and accuracy.Results observed from empiricalstudy, encouraged the further research work in numerous ways like utilization of the proposed method in various approaches, ways to innovate additional effective models to find the proximityvalues for attributes which comprise multiple values. In addition to that the outcomes from the experimental analysis are believed to be encouraging the research to deploy in wide directions including usage of proximity measure in different applications.

References:

1. J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann, 2011.

2. 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.

3. Fuyuan Cao , 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.

4. Wenhao Shu and Wenbin Qian, “Mutual Information-based Feature Selection from Set-valued Data”, 2014 IEEE 26th International Conference on Tools with Artificial Intelligence.

5. S. Guha, R. Rastogi, and K. Shim, “Rock: A robust clustering algorithmfor categorical attributes,” in *Proc. 15th Int. Conf. Data Eng.*, Sydney,NSW, Australia, Mar. 1999, pp. 512–521.

6. 6 .F. Giannotti, C. Gozzi, and G. Manco, “Clustering transactional data,”in*Principles of Data Mining and Knowledge Discovery* (Lecture Notesin Artificial Intelligence), vol. 2431, T. Elomaa*et al.*, Eds. 2002,pp. 175–187.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

7. Joshi, A.; Kaur, R.: A review: comparative study of various clusteringtechniques in data mining. Int. J. Adv. Res. Comput. Sci.Softw. Eng. **3**(3), 67–70 (2013)

8. Ghosh, S.; Dubey, S.K.: Comparative analysis of k-means andfuzzy c-means algorithms. Int. J. Adv. Comput. Sci. Appl.(IJACSA) **4**(4), 35–39 (2013).

9. Velmurugan, T.: Performance based analysis between *k*-means9and fuzzy *C*-means clustering algorithms for connection orientedtelecommunication data. Appl. Soft Comput. **19**, 134–46 (2014).

10. Wang, C.;Dong, X.;Zhou, F.;Cao,L.; Chi, C.-H.: Coupled attributesimilarity learning on categorical data. IEEE Trans. Neural Netw.Learn. Syst. **26**(4), 781–97 (2015).

11. Mukhopadhyay, A.; Maulik, U.; Bandyopadhyay, S.; Coello,C.A.C.: Survey of multiobjective evolutionary algorithms for datamining: part II. IEEE Trans. Evolut. Comput. **18**(1), 20–35 (2014)

12. Y.-I. Kim, D.-W. Kim, D. Lee, and K. H. Lee, "A cluster validation index for GK cluster analysis based on relative degree of sharing," *Information Sciences*, vol. 168, no. 1–4, pp. 225–242, 2004.

13. M. Zhang, W. Zhang, H. Sicotte, and P. Yang, "A new validity measure for a correlation-based fuzzy c-means clustering algorithm," in *Proceedings of the 31st IEEE Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC '09)*, pp. 3865–3868, Minneapolis, Minn, USA, September 2009.

14. Ahmed MN, Yamany SM, Mohamed N, Farag AA, Moriarty T. A modified fuzzy c-means algorithm for bias field estimation and segmentation of MRI data. IEEE Trans Med Imaging 2002; 21:193–9.

15. B. Rezaee, "A cluster validity index for fuzzy clustering," *Fuzzy Sets and Systems*, vol. 161, no. 23, pp. 3014–3025, 2010.

16. Bouveyron, C. and Brunet-Saumard, C., 2014. Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis*, 71, (March 2014) pp.52-78.

17 Banka, H. and Dara, S., 2015. A Hamming distance based binary particle swarm optimization (HDBPSO) algorithm for high dimensional feature selection, classification and validation. *Pattern Recognition Letters*, 52, (2015), pp.94-100.

18. Hedjazi, L., Aguilar-Martin, J., Le Lann, M.V. and Kempowsky-Hamon, T., 2015. Membership-margin based feature selection for mixed type and high-dimensional data: Theory and applications. *Information Sciences*, 322, (2015), 174-196.

19 Duda, R. H. (2001). Pattern Classification and Scene Analysis. *A Wiley-Interscience Publication, New York: Wiley,*.

20. Džeroski, S. (2003). Multi-relational data mining: an introduction. *ACM SIGKDD Explorations Newsletter, 5.1*, 1-16.

21. Aha, D. W. (1992). Tolerating noisy, irrelevant and novel attributes in instance-

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

based learning algorithms. *International Journal of Man-Machine Studies, 36.2* , 267-287.

22.Deza, M. M. (2009). Encyclopedia of distances. *Encyclopedia of Distances, Springer Berlin Heidelberg*, 1-583.

23. Kalousis, A. A. (2006). A unifying framework for relational distance-based learning founded on relational algebra. *Technical Report, Computer Science Department, University of Geneva*.

24. Duda, R. H. (2001). Pattern Classification and Scene Analysis. *A Wiley-Interscience Publication, New York: Wiley,*.

25. Džeroski, S. (2003). Multi-relational data mining: an introduction. *ACM SIGKDD Explorations Newsletter, 5.1*, 1-16.

26. Chen, Y.-L. C.-L.-C. (2003). Constructing a multi-valued and multi-labeled decision tree. *Expert Systems with Applications, 25.2*, 199-209.

27. Chou, S. a.-L. (2005). MMDT: a multi-valued and multi-labeled decision tree classifier for data mining. *Expert Systems with Applications, 28.4* , 799-812.

28. K. LNC Prakash, "Optimal Feature Selection for Multivalued Attributes Using Transaction Weights as

Utility Scale", Proceedings of the Second International Conference on Computational Intelligence and

Informatics ICCII-2017, pp 533-546.

29. Tasca, M. B. (2013). A relevance measure for multivalued attributes. *Journal of Information and Data Management, 4.3*, 421.

30. LNC.Prakash K, "clustering multivalued attribute data using transaction weights as utility scale",
Journal of Advanced Research in Dynamical and Control Systems Vol. 9. Sp– 18 / 2017.

31.
https://relational.fit.cvut.cz/dataset/CORA

32.Van Rijsbergen, C. J. "Foundation of Evaluation." Journal of Documentation 30.4 (1974): 365-
73.
33.Larsen, Bjornar, and Chinatsu Aone. "Fast and Effective Text Mining Using Linear-time Document
Clustering." (1999).
34.Steinbach, M., G. Karypis, and V. Kumar. "A comparison of document clustering techniques." (2000).

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731