Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 3818107, 9 pages
https://doi.org/10.1155/2022/3818107

Research Article

Fuzzy-Based Approach for Clustering Data with Multivalued Features

L. N. C. Prakash K,[1] **M. Vimaladevi**,[2] **V. Deeban Chakravarthy**,[3] **G. Surya Narayana**,[4] **and Asadi Srinivasulu**[5]

[1]*Department of Computer Science and Engineering, CVR College of Engineering, Mangalpalli Hyderabad, India*
[2]*Department of Computational Intelligence, School of Computing, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India*
[3]*Department of Computing Technology, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India*
[4]*Department of Computer Science and Engineering, Vardhaman College of Engineering, Hyderabad, Telangana, India*
[5]*BlueCrest University, Liberia*

Correspondence should be addressed to Asadi Srinivasulu; head.research@bluecrestcollege.com

Received 4 April 2022; Revised 30 April 2022; Accepted 6 May 2022; Published 23 May 2022

Academic Editor: Kuruva Lakshmanna

In analysis of data, objects have mostly been characterized by a set of characteristics known as attributes, which together contained only one value for each object. Besides that, a few attributes in reality could include with more than a single value; such as from a human beside multiple profession characterizations, practises, communication methods, and capabilities, in addition to shipping addresses, of that kind of attributes are referred to as multivalued attributes and are typically regarded as null attributes when data is processed employing machine learning procedures. Throughout this article, another similarity mechanism is introduced that is defined around including multivalued characteristics which can be used for grouping. We propose a model to analyse each factor's relative prominence for different data collection challenges in order to enable the selection among the most suited multivalued elements. The suggested methodology is a clustering technique for development and evolution that employs fuzzy c-means clustering and retains the new and more effective membership component by implementing the proposed similarity metric. Clustering of multivalued variables using fuzzy c-means is the efficient grouping criteria that results; any methodology to group-related data appears viable. The results show that our assessment not only improves previous segmentation methods on the multivalued cluster-based architecture but also helps in the improvement of the standard similarity metrics.

1. Introduction

Clustering is an unsupervised knowledge extraction technique for discovering and organizing related data elements into different groups in massive datasets. Clustering (also known as cluster analysis) is a technique for grouping items into similar patterns which are easier to comprehend and handle. Clustering can be easily performed by a process called k-means method [1–4] that is quite so useful for the ability to cluster big data sets efficiently. Ruspini [5] and Bezdek [6] report fuzzy variants of the k-means approach, in which each characteristic is permitted to have membership functions to every group instead of possessing a definite membership to one group. Furthermore, because these k-means-type methods can indeed work data with numerical information, they can be used very less in fields like data where big categorical or multivalued data sets are widespread. Hierarchical clustering techniques that are using Gower's likeness coefficient [7] or another divergence procedures [8], the PAM computation, and fuzzy-statistical procedures, combined with theoretical clustering techniques are some other techniques for cluster analysis with categorical data. When implemented to enormous categorical-only data sets, all of these techniques suffer from a widely known inefficiency issue.

A group of n items $X = \{X_1, X_2, ,\cdots, X_n\}$ with a collection of s features $A = \{F_1, F_2, ,\cdots, F_s\}$ are used to represent the data for information analysis. The data collection X is characterized therefore in model as a tabular database with n rows along with s dimensions; everywhere each row represents a particular item, as well as each dimension represents an attribute where values with an element are a unique value. In real-world situations, characteristics in a database may have multiple entries for an object in some attributes, like as a human with various employment positions, pursuits, and talents. In questionnaires, banking, schooling, telecommunication services, retail, and medical databases, all of these entries are common. Table 1 shows the data representation of such multiple entries for an object in some attributes.

The data in Table 1 could be described in the subsequent way. Let $\mathbf{X} = \{X_1, X_2, \cdots, X_n\}$ take place the repository, in which every X_i is an object, additionally known as an entity or tuple in the data file, but every row is characterized with a collection of s characteristics $\{A_1, A_2, \cdots, A_s\}$, in which every characteristic A_j might remain with a one or multiple values. If the variable A_i is a single-valued characteristic, then each row X_i in the database X has always single value; however, for each multivalued feature A_j, there seems to be a non-empty collection of values meant for all tuples X_j appearing in the database X. There seems to be two possible ways to accomplish the task of multivalued attribute such as attempting something a little clumsy by putting each value in a new row which increases the size of the database; another one is to create new columns and allow different values to be assigned to it. If there have too many fields of this type, this could result in increasing a significant number of given variables.

The present article includes a fuzzy c-means procedure that extends the earlier work in [9]; it is accomplished through creation of a different methodology for generating the fuzzy fragmentation matrix from multivalued data well within approach of the fuzzy c-means technique [6]. The focus of the present document is a technique for locating fuzzy cluster modes when the simple comparison proximity measure is used for multivalued objects. The k-means algorithm's fuzzy model improves the procedure by allocating strength to items in distinct groups. These optimism values would be utilized to determine cluster core and border objects, supplying more helpful info for attempting to deal with learning objects.

The rest of this paper is arranged as go along: in Section 2, it brings a few of the literary works that will be used throughout the whole paper. Section 3 describes the suggested similarity metric and FCM method. Section 4 describes proposed method. Section 5 provides exploratory evidence demonstrating the efficiency of the planned technique. In the end, Section 6 brings the paper to the conclusion.

2. Relevant Work

Giannotti et al. [10] aimed a grouping approach for data points with transnational records by employing k-means methodology and the Jaccard distance measures towards group the multivalued dataset; however, somehow the technique has a low consolidation. Shu and Qian [11], on the unlabeled items, proposed a measure of similarity. Then, to enable faster the selection process of attributes, a feature-based approach is devised and characterized by mutual information that is included in a decreasing environment. In this study, Giannotti et al. [10] provide a paradigm for separating as well as handling data, i.e., the modelling of discrete data with changeable volume. The authors finally reshape the cluster centroid notion by adapting the precise mathematical segregation conception provided in the k-means approach to represent transaction proximity. In comparison, Ghosh and Dubey's k-means and fuzzy c-means [12] cluster algorithms are based on its efficiency in choosing the optimum data evaluation technique. This clustering technique collected the information into account in the form of locations across various intermediate data objects. FCM is an unsupervised categorization procedure that is being used and employed around a variety of disciplines, including agriculture, scientific, pharmacological, ecologic, medical image processing, categorization, and clustering. The efficiency of the FCM's clustering approaches is compared to those from the k-means grouping approach in this study. Mukhopadhyay et al. [13] offered several multiobjective evolutionary procedures. Each and every time the number of attributes remains large, the binary coding scheme's key limitation cannot be grouped. Furthermore, this study looked at two different methods using multiobjective workable clustering algorithm, MODENAR and MODE, as well as three distinct kinds of information, which is like qualitative, quantitative, and fuzzy data gathering procedures. The experiment revealed that categorical data clustering algorithms may efficiently group large amounts of data with a diverse set of characteristics.

Hedjazi et al. [14] introduced a different feature extraction technique which covers all blending kinds and greater dimensions available facts on participation limitations to enhance the efficiency of fuzzy classifiers. The findings demonstrate that the strategy improves the classification efficiency of both fuzzy classification in addition other related classifiers significantly. According to Zhen, in [15], an instructional ability assessment technique centered on big data fuzzy k-means grouping and information fusion is suggested to fix the challenges of erroneous classification of big data knowledge in standard English training capacity assessment techniques. To begin, the researcher applies the knowledge of k-means gathering to the gathered existing error data and eliminates the data that the procedure recognises uncertain, uses the remainder valid measurements to compute the weighting factor of the altered fuzzy logic algorithm, analyses the weighted average with the node data measured, and obtains the final fusion value. Furthermore, the author combines big data content fusion and the k-means clustering approach, resulting in grouping and index variable integration.

Many generalised FKM techniques have really been developed to this purpose, utilizing breakthroughs in various machine learning approaches. Some authors, for example, sought to alleviate the noise sensitivity problem by

Table 1: Data with multivalued attributes.

ID	Name	Gender	Age	Fluency	Activities
1	Raju	M	29	{Telugu, Urdu}	{Swimming, reading}
2	Rani	F	32	{Telugu, Kannada}	{Music, watching movie}
:	::::::::::::	:::::::::	:::::::	:::::::::::::::::::::::::	::::::::::::::::::::::::::::::::::
n	Jill	M	37	{Kannada, Tamil}	{Reading, sports}

incorporating outlier groups into the FKM design [16]. Ménard et al. suggested a fuzzy generalised k-means process [17] to develop a near relationship among both the framework MEC [16] and FKM. To produce spatial-smooth membership function parameters, Pham introduced a new penalty name to the fitness function of FKM [18]. Cai et al. suggested a quick and vigorous FKM approach to picture separation by incorporating local spatial and grey features [19].

Guo et al. introduced a unique grouping approach in which all the L_{21} norm is utilized to diminish the power of outliers by combining fuzzy k-means and nonnegative spectral grouping into a common structure [20]. Zhang et al. introduced a resilient integrated deep k-means clustering method to give an expressive association expression among observations for deep neural networks, and the norm system of measurement is being applied to regulate the feature function procedure of the auto encoder system [21].

2.1. Data Based on Attributes with Multiple Values. Distance measurements are being used to compare the affinities of two things. The Euclidean distance [22] has been a few of the well commonly employed distance measure when referring to quantitative data in specific classification systems. When dealing with nominal data, distance is usually calculated by assigning a 0 to distinct kinds of values and then a 1 to entirely equal values. When single and multivalued characteristics are present, a unique proximity metric idea must be established that is capable of reliably contrasting different sets. As a result, the researchers propose various criteria for describing the closeness of two sets. The research [23–25] reflects the closeness measure among different features and has been assembled for the implementation of such evaluations in this suggested study.

Distance guesstimate of all research findings produced comparable results; hence, the average similarity test findings were mentioned in this manuscript. A differential evolution-based multivalued attribute data (DEC-MVA) grouping technique, designed by LNC Prakash [26], was adopted to measure the relative relevance of every component in respect to multiple data gathering challenges to support the most effective multivalued characteristics. This approach also created an evolutionary technique that integrates the transaction utility as an optimized process and uses a differential evolution approach. The article's insight offers a novel distance function that suits multivalued properties of multiple types of frameworks; this scale is applicable for both supervised and unsupervised machine learning techniques of data mining research [27]. In almost the same manner [28], RMULT, a multifunctional feature significance test, was utilized for evaluating the importance for classification, along with its multivalued characteristic. This measure is used to evaluate the extent of multivalued categorization features. Nonetheless, because multivalued characteristics combine several quantities, variations of these features correlate to distinct categories [29]. In [30], it is studied on multivalued data and developed techniques to filter out strong however uninteresting rules in association rule mining. Two theories, namely, MMC and MMDT, were explained for multivalued databases in [31, 32]; the two approaches are established by the decision tree methodology. The revised version of MMC is MMDT; of these, the MMC distinguishes features, whereas the MMDT method in addition enhances certain features, to ensure the highest effectiveness of classification details. The study [33] explains a new process to select the best set of values for multivalued features, which makes it simpler to measure their importance for extraction method. This model recommended to choose values built on associated transaction weight, in difference to the general trend of selecting values for multivalued features varying on the frequency. The established model is produced by the utility analysis methods, in which the values are selected corresponding to their importance as a replacement of its existence.

3. Similarity Measure

Basis of the discussion of distance measures in the previous section, important aspects must be considered while selecting a fair distance measure for multivalued grouping. These factors typically would include the sort of analysis and the objective of the research, both of which influence the type of distance measure that will be employed. When determining likeness for multivalued attributes, the occurrence of exact comparable patterns is indeed required, but it is additionally necessary to include limited similarity in the same way as mismatched values of the multivalued feature principles [34]. During grouping, the approach for finding correlation among objects with multiple values is based on multivalued characteristics [35]. Now compared with existing metrics, it permits for the exploitation of several points of comparison to determine grouping similarity. The following is how the similarity of things is determined in this exploration.

$\mathrm{PMA}(X, Y)$ is a similarity computation between two multivalued characteristic values where $X = \{x_1, x_2, x_3, \cdots, x_n\}$, $n \geq 2$ and $Y = \{y_1, y_2, y_3, \cdots, y_m\}$, $m \geq 2$, which is determined by thinking about the closeness among the item values of the multivalued attribute by using Tversky measure.

$$\mathrm{PMA}(X, Y) = \frac{|X \cap Y|}{|X \cap Y| + \beta(\alpha a + (1 - \alpha)b)}, \tag{1}$$

where a and b have being specified by $a = \mathrm{Min}(|X - Y|, |Y - X|)$ and $b = \mathrm{Max}(|X - Y|, |Y - X|)$. The symmetric difference between multivalued attribute values is given in the subsequent way, $X - Y = \{x_i \epsilon X / x_i \notin Y\}$ and $Y - X = \{y_i \epsilon Y / y_i \notin X\}$. Further, $\alpha, \beta \geq 0$ are parameters of the above similarity measure. Setting $\alpha = \beta = 1$ produces the Tanimoto coefficient; setting $\alpha = \beta = 0.5$ produces the Dice coefficient. By considering X to be the prototype and Y to be the variant, then α corresponds to the weight of the prototype, and β corresponds to the weight of the difference. Tversky measures with $\alpha + \beta = 1$ are of particular interest. This formulation also rearranges parameters α and β. Thus, α controls the balance between $|X - Y|$ and $|Y - X|$ in the denominator. Similarly, β monitors the impact of the symmetric variation $|X - Y|$ and $|Y - X|$ versus $|X \bigcap Y|$ in the denominator.

The proximity $\mathrm{PROX}(T_i, T_k)$ between two data vectors T_i and T_k, that are represented by a collection of s quantity of characteristics, is calculated in expressions of specific dimensions' likeness. PMA can be used to assess dimensional similarity for multivalued attributes (X, Y). The subsequent formula can be applied to measure the value of $\mathrm{PROX}(T_i, T_k)$.

$$\mathrm{PROX}(T_i, T_k) = \frac{\sum_{j=1}^{s} \mathrm{PMA}\left(T_i^j, T_k^j\right)}{s}. \tag{2}$$

4. Clustering by Fuzzy C-Means Procedure

The fuzzy c-means methodology addressing multivalued data points is mainly described in this segment. Let $X = \{X_1, X_2, \cdots, X_n\}$ is a data base with n records along with multivalued attributes. Assume that data X_i $(1 \leq i \leq n)$ is characterized by means of a set of attributes $\{A_1, A_2, A_3, \cdots, A_s\}$; here, each attribute A_l is a feature with unique value or with multiple values. Each A_l illustrates a set different value called domain which is symbolized by $D(A_l) = \{a_l^1, a_l^2, \cdots \cdots .a_l^p\}$, where p is the quantity of distinct values of the feature A_l for $1 \leq l \leq s$. When A_l is an attribute with unique value, then each a_l^i $(1 \leq i \leq p)$ is believed as a set with unique value, and when A_l is an attribute with multiple values, then each a_l^i $(1 \leq i \leq p)$ is taken as a set with more than one value, and also $D(A_l)$ is characterized as a predetermined and with order less. Consider X_j be symbolized by $\{x_{j,1}, x_{j,2}, \cdots, x_{j,s}\}$; therefore, X_j be able to reasonably exemplified as a combination of attributes. The intention of the FCM procedure for multivalued data set is to group the data set X hooked on c clusters before diminishing the equation as presented in $L_m(V, C : X)$.

$$L_m(V, C : X) = \sum_{j=1}^{C} \sum_{i=1}^{n} \left(\mu_{ij}\right)^m d_{ij}^{\,2}, \tag{3}$$

where n is the tuples appearing in X; C is the total number of clusters going to form; U is the matrix of membership function, and the components of U are (μ_{ij}); μ_{ij} is the association membership for which the ith tuple belonging to the jth group; d_{ij} is the distance from X_i to C_j, in which C_j denotes the cluster center of the jth cluster; and m is the proponent on μ_{ij} that monitors fuzziness or extent of groups intersect.

The fuzzy c-means methodology concentrates on diminishing L_m with respect to the subsequent limitations on U:

$$\begin{gathered} \mu_{ij} \epsilon [0, 1]; \quad 1 \leq j \leq c; \quad 1 \leq i \leq n, \\ \sum_{j=1}^{k} \mu_{ij} = 1, i = 1 \cdots ..n, \\ 0 < \sum_{i=1}^{n} \mu_{ij} < n, j = 1 \cdots ..c, \end{gathered} \tag{4}$$

where μ_{ij} is the participation level of the record X_i to jth cluster and is furthermore a component of a $n \times c$ matrix $V = [\mu_{ij}]$. $C = \{C_1, C_2, \cdots, C_c\}$ entails the centroids of the fuzzy groups. The cluster centroid C_j is characterized as $\{C_{j1}, C_{j2}, \cdots, C_{js}\}$, and the value m monitors the fuzziness of association of every record.

To group multivalued records, the fuzzy c-means system continues to group data with multivalued attributes cantered on the fuzzy c-means style process. The way of determining the proximity among a cluster centroid as well as a data point, but also the technique for revising the group centroid through each repetition, is presented initially. The closeness among a centroid Ci and a multivalued piece of data X_j is calculated using the similarity metric, which would be the formula given in (1).

The centroids of each different cluster are revised as the group centroid $C_j = \{C_{j1}, C_{j2}, , \cdots, C_{js}\}$ is given; each $C_{jl} \epsilon C_j$ for $1 \leq l \leq s$ centered on the kind of the attribute. When the numerical attribute A_l is present, the C_{jl} is then updated as presented as follows:

$$C_{jl}{}^{k} = \frac{\sum_{i=1}^{n} \left(\mu_{ij}{}^{(k-1)}\right)^m x_i}{\sum_{i=1}^{n} \left(\mu_{ij}{}^{(k-1)}\right)^m}, \quad j = 1 \cdots C. \tag{5}$$

For the categorical or multivalued attribute A_l, the centroid value C_{jl} is updated as given as follows:

$$\begin{gathered} C_{jl}{}^{k} = a_l{}^{(q)} \epsilon D(A_l), \\ \text{where} \sum_{x_{il} = a_l^{(q)}} \left(\mu_{ij}\right)^m \geq \sum_{x_{il} = a_l^{(t)}} \left(\mu_{ij}\right)^m, 1 \leq t \leq p \text{ and } q \neq t. \end{gathered} \tag{6}$$

To improve the empirical function $L_m(U, C : X)$ discussed in equation (3) with projected centers that exist

Input: Specify the number of clusters C, the membership degree m >1, and the error.
Output: the centres of clusters and the membership degrees μ_{ij}.

step 1. Choose cluster centers randomly c (0).
step 2. Initiate k =1.
step 3. Replicate
step 4. Compute the association matrix V(k) utilizing the centroids c(k − 0).

$$\mu_{ij}(k) = 1/\sum_{z=1}^{c}\left(PROX(C_j, X_i)/PROX(V_z, X_i)\right)^{2/m-1}$$

step 5. Renovate the centroids of fuzzy clusters c(k) utilizing V(k):

$$C(t) = \{C_{j1}, \cdots, C_{jl}, \cdots, C_{js}\} \text{ for i = 1, 2… c. For each} C_{jl} \epsilon A_l.$$

$$V_{il}{}^{t} = \begin{cases} \left(\sum_{i=1}^{n}\left({\mu_{ij}}^{(k-1)}\right)^m x_i / \sum_{i=1}^{n}\left({\mu_{ij}}^{(k-1)}\right)^m\right) \text{ If } A_l \text{ is continuous valued attribute} \\ \begin{cases} {a_l}^{(q)} \epsilon D(A_l) \text{ If } A_l \text{ is categorical valued attribute} \\ \text{where} \sum_{x_{jl}=a_l^{(q)}} (\mu_{ij})^m \geq \sum_{x_{il}=a_l^{(t)}} (\mu_{ij})^m, 1 \leq t \leq p \end{cases} \end{cases}$$

step 6. Compute k=k+1
step 7. If no development in L_m, then stop the procedure, otherwise go to step 4.
step 8. Return ci the centers of clusters and the membership degrees μ_{ij}.

Algorithm 1: The organized clustering technique for data with multiple value attributes.

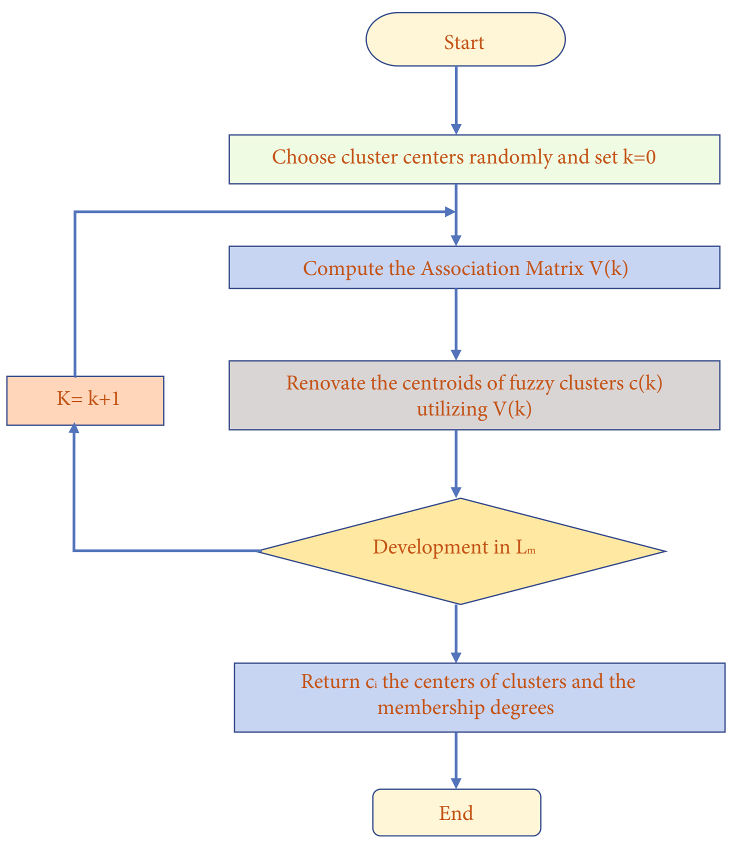

Figure 1: Fuzzy c-means clustering procedure for multivalued data.

defined in equation (5) and equation (6), the intended procedure that applies the fuzzy c-means kind model to group data set with multivalued attributes is given in Algorithm1. Figure 1 represents the flow of the fuzzy c-means procedure for multivalued data.

5. Assessment of Effectiveness and Experimentation

This portion of the article presented epidemiological findings on data sources, assessment methods, and enabling technologies of the developed model. To approximate the usefulness of the suggested clustering procedure, experimentations are accomplished on k-means categorization which aims to cluster the dataset. The distance function [29], the average similarity used here is as given as follows:

$$\mathrm{SIM}(X, Y) = \begin{cases} \dfrac{\sum_{I=1}^{n} \sum_{j=1}^{m} s\left(x_i, y_j\right)}{|X||Y|}, & \text{If } X \neq Y, \\ 0 & \text{If } X = Y, \end{cases} \tag{7}$$

where $d(x_i, y_j)$ is the similarity between every pair of principles constructed from the tuples X and Y that is depicted as presented as follows:

$$s\left(x_i, y_j\right) = \begin{cases} \left|x_i - y_j\right|, & \text{if } x_i \text{and } y_j \text{ are continuious values,} \\ 0 & \text{if } x_i \text{and } y_j \text{ are discrete and } x_i = y_j, \\ 1 & \text{if } x_i \text{and } y_j \text{ are discrete and } x_i \neq y_j. \end{cases} \tag{8}$$

The scripts are written in scripting language and are used to analyse the effectiveness on the resulting clusters. This segment delves into the representation and characteristics of the real dataset used in the investigation. CORA [36] is the real source of data used during the experimental studies.

5.1. Real Database. The CORA [36] dataset is of particular importance in the study because it contains 2,708 data recordings which performs a significant role in investigation. All data tuples are a scholarly way of contributing the one of seven groupings, which include machine learning techniques, CBR designs, probabilistic methodologies, rule-based studying strategies, neural network-based genetic methods, and theory centered designs. Each one data tuple contains several items with 1,433 different words known as attributes. Citing and cited publications are the value sets of almost any two features that can carry multiple values. Each CORA article contains a subset of 5,429 special example characteristics selected as a group of multiple values for such characteristics typically require several values. The correctness in addition to standard of work methodology is ascertained by incorporating different group persistence specifications such as cluster pureness and cluster HM, as well as contrasting constructs of both. To accomplish this, the recommended data files are chosen as information sources focused on topic viewpoints. Furthermore, categorization of these files into repositories is demonstrated to aid in the best possible perseverance of clusters based on the chosen specifications.

5.2. Evaluation of the Proposed Solution Characteristics and Approaches. Purity is an effective feedback metric of cluster efficiency used in cluster analysis in the metric domain [0..1]; also, this is the percent of the overall quantity of items (data items) that were properly categorised. Purity is a metric for how many clusters include a single class. The following is an example of its calculation: count the number of observations from either the cluster's most common category for each group. The inverted purity metric is used and is necessary for assessing data clusters as comparable categories. This inverted statistic is critical for determining which cluster has the highest recall value for each category. Because this factor is impotent to negate the mixture of numerous records gathered from various groups, determining a cluster containing all the tuples yields maximum amount to inverted purity. In addition to the foregoing two factors, the HM of document clusters is taken into account. The inverse purity and conjunction of purity, referred to as F-measure, are calculated on each category of the cluster with the highest combined precision and recall [37–40]. The procedure was tested on a system with a 4 GB RAM capacity and an i5 CPU. The scripts use the Python programming language to describe how to measure the outcomes on the generated clusters.

Table 2: Underlying data statistics as well as the outcomes achieved.

	Proposed clustering with Tversky measure	k-Means clustering with average similarity measure
The size of the dataset (rows)	2708	2708
Count of single valued attributes	1433	1433
Count of attributes with multiple values	2	2
Classes in the database (clusters)	7	7
Overall average F-measure	0.93	0.89
Overall average cluster purity	0.94	0.91
Overall average clustering accuracy	0.88	0.85

5.3. Study of the Proposed Work from a Statistical and Practical Perspective. The proposed approach aids in the improvement of clusters that are derived from datasets of documents with multiple value attributes for the reason that the F-measure of individual groupings is superior; also, the amount of purity for all discovered clusters observed with higher accuracy percentages. To emphasise the significance of the proposed methodology, the k-means grouping procedure is utilized by using the averaged similarity metric. In addition, the proposed model achieves best possible purity and F-measure characteristics. The parameter estimates that resulted are more impactful than the values that contributed from previous systems. The quantitative information relating to the exploratory assessment of the research solutions is illustrated in Table 2.

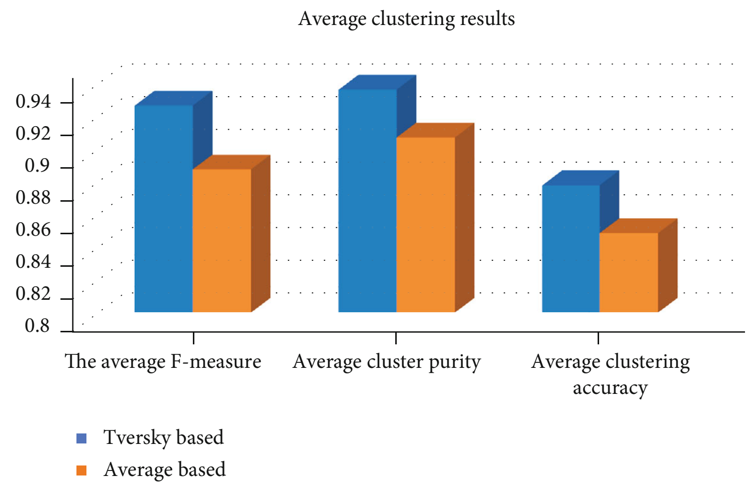

FIGURE 2: Average clustering parameter estimates.

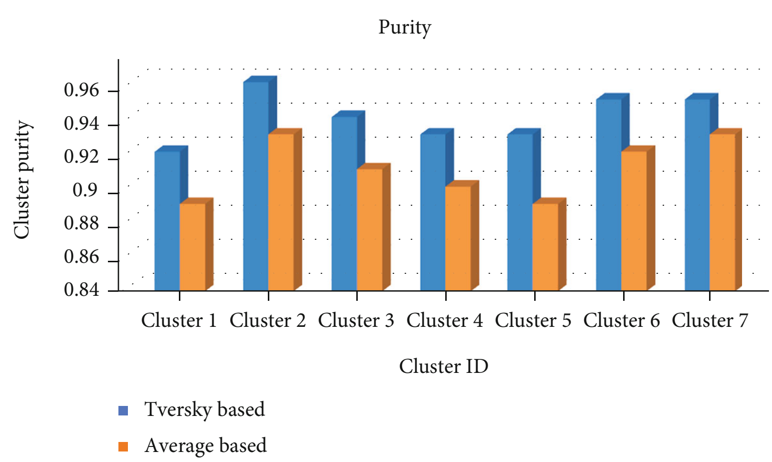

FIGURE 3: Divergent cluster's purity.

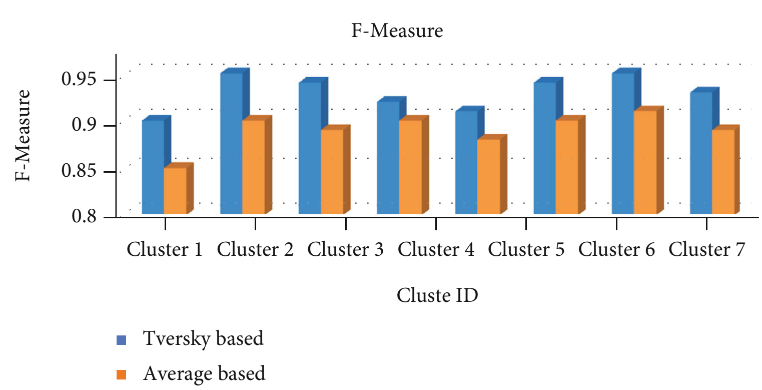

FIGURE 4: Divergent cluster's F-measure.

The given Figure 2 depicts the results of Table 2 which expresses that the proposed measure produces optimal results when compared with existing with respect to the average cluster performance evaluation methods.

The purity, F-measure, and accuracy values of different clusters are shown in the diagrams below.

Figure 3 depicts the purity for both strategies. It signifies the precise terms of percentage value between an assessed cluster's obtained and original true data. Figure 4 illustrates the F-measure for both schemes. It embodies the accurate expressions of percentage value between considered cluster's obtained and original true data. Figure 5 portrays the

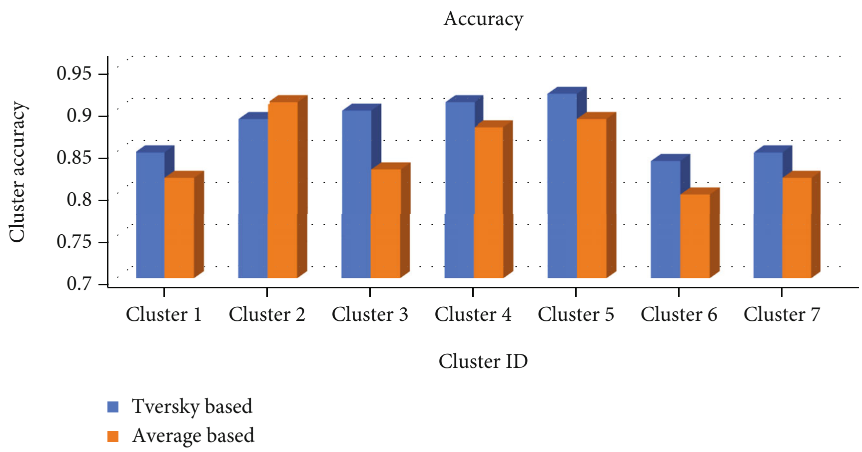

Figure 5: Accuracy values for divergent clusters.

accuracy for both approaches. It signifies the accurate terms of proportion value between an assessed cluster's found and original true data.

6. Conclusions

The research presented in this publication is the step in the direction of clusters with multivalued data. Several analytic techniques demand clustering based on unordered multivalued features. Clustering multivalued data has certain unique issues that does not present in single valued information. This research looked at a comparison measure for data objects with multiple values that are unordered. The investigational findings showed that the anticipated system is appropriate as a clustering technique when applied on the CORA data set [36], which contains both multiple values and single value characteristics.

The research investigation also shown the use of the suggested distance function applied on the given data during the cluster learning process. The developed model's effectiveness was assessed by examining it to the results of another comparable model known as average distance with respect to purity, F-measure, and accuracy and produced significant results. The results of the experimental study stimulated further research in different kinds of ways, including the use of the suggested method in different methods and ways to develop new useful models for determining similarity of the characteristics with multiple values. Additionally, in future, the conclusions of the practical assessment are influencing the investigation to develop in a diversity of ways, containing the use of proximity measures in other applications such as analysis of surveys, reviews, investigations, and medical data analysis in which the multivalued data is involved.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request (head.research@bluecrestcollege.com).

Conflicts of Interest

The authors declare that they have no conflicts of interest to report regarding the present study.

Authors' Contributions

KLNC Prakash contributed to the conceptualization, data curation, formal analysis, methodology, software, and writing—original draft. M. Vimaladevi contributed to the supervision, writing—review and editing, project administration, and visualization. V. Deeban Chakravarthy contributed to the visualization, investigation, formal analysis, and software. G. Surya Narayana contributed to the data curation, investigation, resources, and software. Asadi Srinivasulu contributed to the supervision, writing—review and editing, visualization, and methodology.

References

[1] M. R. Anderberg, *Cluster Analysis for Applications: Probability and Mathematical Statistics: A Series of Monographs and Textbooks*, vol. 19, Academic press, 2014.

[2] G. H. Ball and D. J. Hall, "A clustering technique for summarizing multivariate data," *Behavioral Science*, vol. 12, no. 2, pp. 153–155, 1967.

[3] J. Ak and R. C. Dubes, *Algorithms for Clustering Data. Englewood Cliff*, Prentice Hall, 1988.

[4] J. B. Mac Queen, "Some methods for classification and analysis of multivariate observations," in , Article ID 669871*Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 281–297, Berkeley, CA, 1967.

[5] E. R. Ruspini, "A new approach to clustering," *Information and Control*, vol. 15, no. 1, pp. 22–32, 1969.

[6] J. C. Bezdek, "A convergence theorem for the fuzzy ISODATA clustering algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-2, no. 1, pp. 1–8, 1980.

[7] J. C. Gower, "A general coefficient of similarity and some of its properties," *Biometrics*, vol. 27, no. 4, pp. 857–874, 1971.

[8] K. C. Gowda and E. Diday, "Symbolic clustering using a new dissimilarity measure," *Pattern Recognition*, vol. 24, no. 6, pp. 567–578, 1991.

[9] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data Mining Knowledge Discovery*, vol. 2, no. 3, pp. 283–304, 1998.

[10] 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, pp. 175–187, Springer, Berlin, Heidelberg, 2002.

[11] W. Shu and W. Qian, "Mutual information-based feature selection from set-valued data," in *2014 IEEE 26th International Conference on Tools with Artificial Intelligence*, Limassol, Cyprus, 2014.

[12] S. Ghosh and S. K. Dubey, "Comparative analysis of k-means and fuzzy c-means algorithms," *International Journal of Advanced Computer Science and Applications*, vol. 4, no. 4, pp. 35–39, 2013.

[13] A. Mukhopadhyay, U. Maulik, S. Bandyopadhyay, and C. A. C. Coello, "Survey of multiobjective evolutionary algorithms for data mining: part II," *IEEE Transactions on Evolutionary Computation*, vol. 18, no. 1, pp. 20–35, 2013.

[14] L. Hedjazi, J. Aguilar-Martin, M. V. Le Lann, and T. Kempowsky-Hamon, "Membership-margin based feature selection for mixed type and high-dimensional data: theory and applications," *Information Sciences*, vol. 322, no. 2015, pp. 174–196, 2015.

[15] C. Zhen, "Using big data fuzzy K-means clustering and information fusion algorithm in English teaching ability evaluation," *Complexity*, vol. 2021, Article ID 5554444, 9 pages, 2021.

[16] D. Tran and M. Wagner, "Fuzzy entropy clustering," in *Ninth IEEE International Conference on Fuzzy Systems. FUZZ- IEEE 2000 (Cat. No.00CH37063)*, vol. 1, pp. 152–157, San Antonio, TX, USA, 2000.

[17] M. Ménard, V. Courboulay, and P.-A. Dardignac, "Possibilistic and probabilistic fuzzy clustering: unification within the framework of the non-extensive thermostatistics," *Pattern Recognition*, vol. 36, no. 6, pp. 1325–1342, 2003.

[18] D. L. Pham, "Fuzzy clustering with spatial constraints," in *Proceedings. International Conference on Image Processing*, vol. 2, Rochester, NY, USA, 2002.

[19] W. Cai, S. Chen, and D. Zhang, "Fast and robust fuzzy *c*-means clustering algorithms incorporating local information for image segmentation," *Pattern Recognition*, vol. 40, no. 3, pp. 825–838, 2007.

[20] M. Guo, R. Zhang, F. Nie, and X. Li, "Embedding fuzzy k-means with nonnegative spectral clustering via incorporating side information," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp. 1567–1570, Torino, Italy, 2018.

[21] R. Zhang, H. Tong, Y. Xia, and Y. Zhu, "Robust embedded deep k-means clustering," in *Proceedings of the 28th ACM international conference on information and knowledge management*, pp. 1181–1190, Beijing, China, 2019.

[22] M. M. Deza, "Encyclopedia of Distances," in *Encyclopedia of Distances*, vol. 1-583, Springer, Berlin Heidelberg, 2009.

[23] A. A. Kalousis, *A Unifying Framework for Relational Distance-Based Learning Founded on Relational Algebra*, Computer Science Department, University of Geneva, 2006, Technical Report.

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

[25] S. Džeroski, "Multi-relational data mining," *ACM SIGKDD Explorations Newsletter*, vol. 5, no. 1, pp. 1–16, 2003.

[26] L. N. C. Prakash K, "Clustering multivalued attribute data using transaction weights as utility scale," *Journal of Advanced Research in Dynamical and Control Systems*, vol. 9, 2017.

[27] A. Balakrishnan, K. Ramana, G. Dhiman et al., "Multimedia concepts on object detection and recognition with F1 car simulation using convolutional layers," *Wireless Communications and Mobile Computing*, vol. 2021, Article ID 5543720, 21 pages, 2021.

[28] M. B. Tasca, "A relevance measure for multivalued attributes," *Journal of Information and Data Management*, vol. 4, no. 3, p. 421, 2013.

[29] K. Ramana, M. Ponnavaikko, and A. Subramanyam, "A global dispatcher load balancing (GLDB) approach for a web server cluster," in *International Conference on Communications and Cyber Physical Engineering 2018*, pp. 341–357, Singapore, 2018.

[30] F. Min and W. Zhu, "Granular association rules for multi-valued data," in *2013 26th IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, Regina, SK, Canada, 2013.

[31] Y. L. Chen, C. L. Hsu, and S. C. Chou, "Constructing a multi-valued and multi-labeled decision tree," *Expert Systems with Applications*, vol. 25, no. 2, pp. 199–209, 2003.

[32] S. Chou and C. L. Hsu, "MMDT: a multi-valued and multi-labeled decision tree classifier for data mining," *Expert Systems with Applications*, vol. 28, no. 4, pp. 799–812, 2005.

[33] K. L. N. C. Prakash, "Optimal feature selection for multivalued attributes using transaction weights as utility scale," in *Proceedings of the Second International Conference on Computational Intelligence and Informatics ICCII-2017*, pp. 533–546, Singapore, 2017.

[34] P. G. Krishna and D. L. Bhaskari, "An efficient fuzzy c-means clustering algorithm for multi-valued data sets," *Information Technology in Industry (ITII)*, vol. 9, no. 1, 2021.

[35] D. S. Prashanth, R. Mehta, K. Ramana, and V. Bhaskar, "Handwritten Devanagari Character Recognition using modified Lenet and Alexnet convolution neural networks," *Wireless Personal Communications*, vol. 122, no. 1, pp. 349–378, 2022.

[36] https://relational.fit.cvut.cz/dataset/CORA.

[37] K. Ramana, T. Krishna, C. Narayana, and M. P. Kumar, "Comparative analysis on cloud computing and service oriented architecture," *International Journal of Advanced Research In Technology*, vol. 1, no. 1, pp. 22–28, 2011.

[38] C. J. Van Rijsbergen, "Foundation of evaluation," *Journal of Documentation*, vol. 30, no. 4, pp. 365–373, 1974.

[39] B. Larsen and C. Aone, "Fast and effective text mining using linear-time document clustering," in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 16–22, San Diego California USA, 1999.

[40] M. Steinbach, G. Karypis, and V. Kumar, *A Comparison of Document Clustering Techniques*, 2000.