International Journal of Communication Networks and Information Security

ISSN: 2073-607X, 2076-0930

*Volume **14** Issue **1s** Year **2022** Page **79:90***

Hybrid Multi Attribute Relation Method for Document Clustering for Information Mining

S. Tejasree[1], Dr. B. ChandraMohan[2*]

[1]Research Scholar, School of Computer Science and Engineering (SCOPE), VIT University, Vellore, Tamilnadu, India
Email: samakoti.tejasree2016@vitstudent.ac.in

[2]Associate Professor, School of Computer Science and Engineering (SCOPE), VIT University, Vellore, Tamilnadu, India.*
Email: chandramohan.b@vit.ac.in
Correspondence Email: chandramohan.b@vit.ac.in

Article History

Received: 13 July 2022
Revised: 20 September 2022
Accepted: 26 October 2022

Abstract

Text clustering has been widely utilized with the aim of partitioning specific documents' collection into different subsets using homogeneity/heterogeneity criteria. It has also become a very complicated area of research, including pattern recognition, information retrieval, and text mining. In the applications of enterprises, information mining faces challenges due to the complex distribution of data by an enormous number of different sources. Most of these information sources are from different domains which create difficulties in identifying the relationships among the information. In this case, a single method for clustering limits related information, while enhancing computational overheads andss processing times. Hence, identifying suitable clustering models for unsupervised learning is a challenge, specifically in the case of Multiple Attributes in data distributions. In recent works attribute relation-based solutions are given significant importance to suggest the document clustering. To enhance further, in this paper, Hybrid Multi Attribute Relation Methods (HMARs) are presented for attribute selections and relation analyses of co-clustering of datasets. The proposed HMARs allow analysis of distributed attributes in documents in the form of probabilistic attribute relations using modified Bayesian mechanisms. It also provides solutions for identifying most related attribute model for the multiple attribute documents clustering accurately. An experimental evaluation is performed to evaluate the clustering purity and normalization of the information utilizing UCI Data repository which shows 25% better when compared with the previous techniques.

Keywords: Data Mining, Clustering, Hybrid Multi Attribute Relation Methods (HMARs), documents, information mining, UCI.

1. Introduction

Wide range of domain data are getting gathered at breakneck speeds and need to be aggregated across storages making it imperative for strategies and computational tools to analyse relationships between stored document attributes and extract useable information. A few methodologies based on a predictive

network for document clustering have been proposed by Brockmeier[1], semantic short text analysis [2], and a clustering of new products through collaborative decisions [3]. Filtering information from unfiltered document data and converting them into informative knowledge is quite a challenging task. In order to assess their everyday operations, most modern business systems usually incorporate volumes of data records with various attribute sets. Still in real-time analysis of business data, it will be luxurious as data is dispersed across different sources, and aggregating these multiple data sources is difficult due to time constraints.

Clustering is a data categorization strategy that is both accurate and computationally efficient where sizes, volumes, and complexities of classifications should be considered for processing large databases. Several different techniques have been present to facilitate efficient clustering in [4-5]. In general, the data distribution of indeterminate objects can be expressed by the probability of distribution [6-7]. The difficulties of grouping such multilevel objects are being point to by the probability distribution, which happen in several situations. The most effectual methods are to demonstrate this characteristic and the mechanism to select the attributes and through which it can reduce the dimensions of the search for a meaningful attribute in a collection of attributes of objects [8-12].

To govern attribute selection, attribute selection techniques are generally classified as "supervised" or "unsupervised." The approach of choosing supervised attributes [13-16] leads a collection of significant and relevant attributes by using the link between properties and name information. As a result, most studies seek to extract important values for analyses; nevertheless, selecting the right features are difficult processes due to diverse information models making clustering of relevant information complex.

Information mining based on relevant characteristics of text documents is effective techniques to identify document relevance for IRs (information retrievals). Current techniques utilise long-term approaches; languages from trainings are used to convey crucial features, but low-level supports are an issue. Cluster analysis using techniques like "k-means," "k-medoids," have been actively investigated for decades for gathering information effectively, with a special focus n distance-based mass analysis. These strategies are most effective when dealing with data with small number of dimensions, but become computationally expensive or even uncertain in their analyses of multiple attribute data. Therefore, finding collections of data documents in uncertain spaces with multiple attribute values is complex and specifically raises the issues when data is irregular and distorted [17].

This study focuses on unsupervised attribute selection issues that arise as a result of multi-value information that implies attribute choices and unlabelled relation difficulties. The HMARs is described as a method for determining the most suitable attribute for most relevant data clustering which shows improvement in precision, purity etc. The HMARs uses the classic k-means technique to create the most related clustered data based on their multiple attribute mean ratio of similarity. The result uses a Hybrid Multi Attribute Relation Method to create the most influential characteristics that should be considered for clustering improvement. Through analysing the condition dependency among these qualities, it adapted a Bayesian process to compute the probability similarity between the attributes. It will provide the sparsity for the attribute selection and reduce irrelevant attribute from the multiple attribute data to improvise the accurate clustering of the unsupervised data.

The paper's next portion is divided into five sections. The second section delves into similar research in the context of clustering algorithms and attributes selection. The suggested work approach is discussed in section 3, and the experimental results and their analysis are offered in section 4. The final portion of this work addresses the work's conclusion.

2. Related Works

Changing qualities of information descriptions stem from different data formats like papers, photos, healthcare records, and informative mining [18]. Multi dimensional composites of multiple values in many characteristics have increased multi-dimensional data processing for multiple values in many sectors. The bulk of these attributes' values may be unique with some as links between data them along with redundant and noisy information which hinder selections by classic learning models generating flaws like inefficiencies and poor performances. As a result, improving precisions and certainty of results is a very complex issue and it is critical to get rid of unnecessary or repetitive features that are linked while groupings of traits[19].

In particular, the *k*-mean algorithm [20-21], is often used to make object comparison based on simple identical similarity and distance of objects to classification and to rather than identifying the means of group clusters, which items are connected with "mean" The clusters are assigned to a collection in a group of domains which consist of similar or nearby values of the attributes, such as the data point that specifies to an individual attribute. Besides, in conjunction with the *k*-means algorithm, one can even utilize the statically and categorical values to clustering of the large databases from a different domain. A cluster may have one or more modes in the *k*-mean clustering; however, it should be noted that the algorithm relies heavily on choosing the mode of the clustering procedure.

2.1 Importance of Attribute Selection

An important function and a widely utilised method in numerous domains of information analysis and mining is detecting acceptable and successful productivity aspects [22-23]. It was suggested that a method for selecting a multiplicity of features and suggesting a "machine learning application" be devised. According on the use of information labelling techniques, these algorithms may be classified as "supervised algorithms", "semi-supervised algorithms", and "unsupervised algorithms" [24-25]. Supervised approaches allow for the selection of qualities by learning the properties of the many attributes in their category. Various current methodologies, such as sparsity-based methods, can be used to understand the relationship between the qualities. However, the data revealed might be prohibitive or exceedingly difficult for a human to process. Data sets including tiny tags and huge data without tags, on the other hand, may be easily customised to form a broad collection of data. The difficulty is outlined in the "small-label sample problem" [26], which is a challenging challenge for the supervised algorithm to solve.

Because it lacks sufficient information about data labels and tiny supervised algorithms, which mistakenly remove many of the requested characteristics or fail due to an incorrect selection attribute. Hence, "semi-supervised attribute selections" were developed for unlabelled/labelled data as they can lead to finding attributes by selecting unsupervised attributes. This makes it more difficult to assess attributes of interest that keep certain qualities of objects in the absence of relevant labels. This result in the need for proficient developments and highly autonomous techniques for selecting unsupervised qualities which need to handle costs of labelling [27].

2.2 Importance of Attribute Selection in Clustering

Unsupervised clustering suffering from defining attributes accordance to the requirements stated for the most difficult clustering criterion, due to a lack of class/label information that specifies a part of a specific cluster's characteristics [28]. Cluster-based attribute identification techniques utilize functional ideas in generating virtual tags, selection of attributes, and pseudo-tags for data duplication.

The difficulty is exemplified by the representation of an attribute vector, in which each occurrence of the attribute is linked to a set of category labels, is solved by the research and selection of multiple attributes. Current techniques process a collection of symmetric characteristics to learn from multiple attribute data (i.e., the instance illustration of every example used to distinguish all RC labels). This broad technique, however, may not be the ideal option for the answer because each tag has its own unique qualities.

In [29] propose LIFT, a method for learning from multiple attribute data that uses name characteristics to take advantage of differences in distinct category labels. Initial assessments of labels' as positive and negative states in groups are executed and then groups are trained and tested to develop label-specific attributes. Extensive testing on 17 reference data sets proved LIFT's superiority over traditional multi-label learning methods, as well as its named attribute efficiency.

In [30] described the attribute selection function that has been implemented for pseudo-label information that is used to spectrally collect data that are simultaneously executed for all cases. In [31] presented that according to the assumption of the label, it can be expect by a "linear classifier decision" to obtain an instance of data input combining discriminatory investigation with the "$l_{2,1}$ -norm minimization" integrated framework to form a common structure for independent attribute selection. Even, provide an attribute-based assumption to improve cluster quality, and also carry out pseudo-label creation for incorporates distinguish analysis for unsupervised attribute selection.

In [32] suggested novel forces for fuzzy clustering of bulky and multi-dimensional data, which are useful for document categorization. Integrating fuzzy collections especially created for documents into

an effective scheme for addressing a big-scale problem is the core principle behind taking huge and large criteria into consideration when creating challenges. Sampling-Extension, Divide Ensemble and Single-Pass, were found as three exemplary techniques in a fuzzy clustering to handle bulky and scaling data. Various studies on real-time huge databases have used these methods, and the findings suggest that these methods regularly outperform other approaches to document classification.

According to [33], this issue lowers the quality of aggregation findings and suggested unique correlations for promoting the normal matrices by recognizing unusual items in the dataset with group comparisons. An efficient correlation-based approach was offered to validate key similarities. Despite efforts to solve acquisitions classified information across groups and evaluated their findings with traditional algorithms which regrettably create chunks of data based on partial data. Many items in dataset collections were left as anonymous since primary matrix information group only indicated group data relationships. To get the final aggregate results, the study applied the graph splitting technique to weighted double-sided graphs composed of repeating matrices.

Cluster analysis is a methodology of studying data to relate a group of objects into identical groups based on their high similarity of the elements of the group. However, the development of the latest cluster technology is reflected by the natural vector of quantitative properties in the multi-dimensional space of the database, which is usually recorded as a distribution of data weight probabilities [34].

Although various discrepancies and distances have been detected in data collection analysis based on the aims and character of the data, there have been less approaches to evaluate uncertain multiple attribute data. The goal of this study is to suggest techniques for improving information clustering on MA data for use in various information extraction applications.

3. Proposed Methodology

In the analysis of traditional multiple attribute data, distance similarities between data are important. Different distance measures operate depending on the type of measurements chosen. Despite the fact that in conventional data analysis, distances too many similarities are defined by objects and data characteristics, the same is applied is advice for co-clustering determined matrix-based data analysis.

PCAs (Principal Component Analysis) are well-known methods for extracting properties from dependent data with distinct features. As a consequence, the suggested HMARs in this work rely on relationships between data attributes and also offer unique criteria for attribute selection using PCAs. The initial k-cluster results are formed utilizing the standard k-mean approach from the perspective of data relation features, and the acquired results are then utilised to calculate the conditional dependency-based matrix to produce the most desirable clustering attributes. The intended flow diagram is shown in Figure 1.

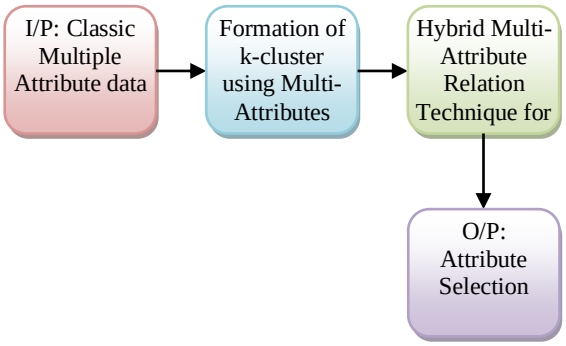


Figure 1. Flow Diagram for Proposed Method

3.1 Creation of k-Cluster using multi-Attributes

The connection matrices are defined by using combined aggregations to represent links between several attributes, where one attribute determines positive effects on other attributes. The set of edges are links between structures formed by term node sets in the graph and pairs of vertices in general. It's worth noting that graphs with form edges necessitate relative edge considerations, and graphs with many clusters inside groups necessitate assignments of point vertices to cluster categories.

To generate a collection of clusters from different attribute datasets, KMCs (k-means clustering) are utilised. They take a series of data records as inputs and perform repeated comparisons based on attribute values until their cluster categories are found. The k value of the approach is predetermined by cluster kind counts. The creation process is depicted in Figure 2.

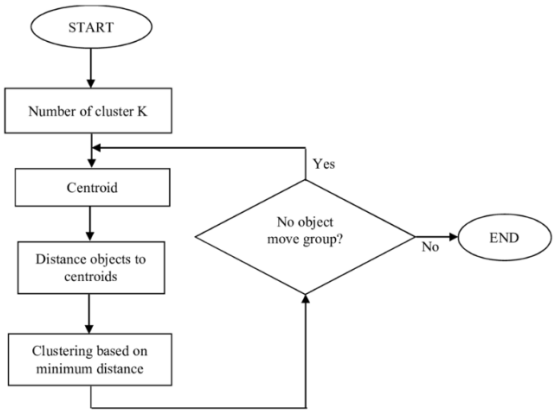


Figure 2. Results created initially using a k- Means Cluster Technique

The aforementioned clustering approach generates sets of data with their important qualities in terms of similarity between attribute values. The closest mean similarity amongst characteristics is used to compute the ratio of attribute similarity between data records. If A and B are two data records with a small number of unique values, then Eq. gives the mean ratio of similarity (1)

$$Min_{sim(A,B)} = \frac{|A_{multi\_att} \cap B_{multi_{att}}|}{|A_{multi_{att}}|.|B_{multi_{att}}|} \quad (1)$$

Since a result of this grouping, there may be a large number of outliers' records, as the mean similarity between these data records may be low, with large disparities in other attribute values. In such circumstances, learning the cluster data is critical in order to facilitate accurate data in the correct cluster. We use a Hybrid Multi Attribute Relation Method for the clustered data to improve the purity of the clustering by computing the Conditional Dependency (CD) of their attributes. The phases that follow will assist us in comprehending how the K-Means clustering approach works:
Algorithm: K-means clustering algorithm
Input: number of clusters K and a collection of objects D.
Output: distribute objects D to clusters K.

1. Generate “k” number of clusters randomly.
2. Calculate the distance between each data points to each of the centres; each data point is assigned to a cluster.
3. The centroids are updated by calculating the mean value of all data points in the respective cluster.
4. Repeat step 2, with respect to new centres.
5. If the assignment of cluster for the data points changes, repeat step 3 else stop the process.

K-means execute the Expectation-Maximization strategy to solve the issues. The Expectation-step is utilized to allocate data points to the nearest cluster, and the Maximization-step is utilized to calculate the centroid of every cluster.

3.2 Hybrid Multi Attribute Relation Method for Clustering

To build an Attribute Relation Matrix (ARM) for the clustered data formed using the *k*-mean technique utilizing attribute value similarities. A set of most frequent and unique attribute of the cluster is build up, where minimum support is to be >=2 for an attribute is consider as *U(A)*.
The ARM operates on a set of cluster dataset created is representing as *Z*. Each instance of *Z*have *S* set of data records, and each instance of *S* have *T* number of attributes, and for each attribute it has *U*

number of distinct values. Now to have a probabilistic relation among each instance of *S* we relate each instance value of a *T* to other instances of *S*AS shown in Fig.3.

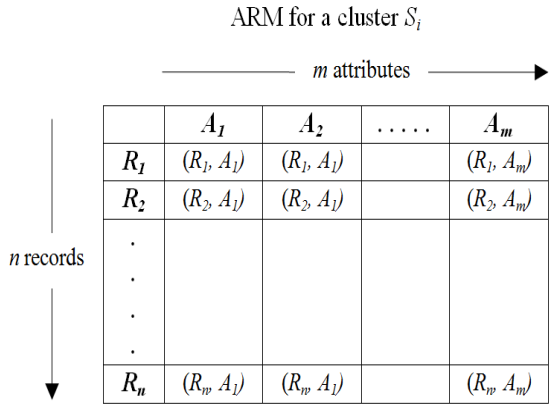


Figure 3. ARM Illustration for a Clustered Set

We looked at two data records that are expected to be adjacent if the probability similarity criterion is met. As a result, the clustering quality is dependent on the threshold modifications being sensitive to support the specified correlation approximations. During attribute generation steps, it may be necessary to set extremely low support thresholds in order to locate collections of extended and overlapping improved characteristics that are appropriate for matching scope of provided test circumstances.
Let's considered that every Cartesian product among R_n and A_m is agreed by a value $V_{(1,\ 1)}$ to $V_{(n,\ m)}$ utilizing the Eq. (2).The product of "*row* × *col*" will be *(0, 1)* in the matrix.

$$p\left(V_{(n,m)}\right) = \prod_i \frac{p(R_n(A_m) \cap U(A))}{p(U(A))} \qquad (2)$$

$$\text{where, } p\left(V_{(n,m)}\right) = \begin{cases} 1, & if \quad R(A) \in U(A) \\ 0, & otherwise \end{cases}$$

Now, to compute the probability of attribute CD with respect to every attribute a Bayesian probability ratio is calculated utilizing the Eq. (3).

$$p\left(CD(A_m)\right) = \frac{\sum_{x=1}^{n} p(V_{(n,m)})}{Number\ of\ R} \qquad (3)$$

As a consequence, the most favoured characteristics are determined for grouping multiple attribute data records while computing attribute CD values using Eq. (3). The characteristic with highest CDs are regarded to have the most clustering-inducing features and can act as focal points with other attributes in relation graphs. Figure 4 shows that A4 is the most important property, while the others are dependent qualities whose inclusion in sequence will improve the purity and NMI of the clustering data.

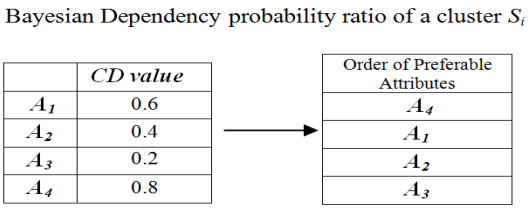


Figure 4. Attributes Preference Order Utilizing CD Value

It will analyse the efficacy of clustering for varied number of attribute selections in multiple attribute datasets based on knowledge of the attribute probability connection. The multi-value data sets clustering will be resolved in the subsequent part based on an order of desirable attributes collecting, and the performance of the recognised techniques [Nguyen et al. (2016)] of clustering approaches will be compared with the suggested methodology.

4. Result and Discussion

4.1 Datasets

For the assessment of the proposed HMARs we download the three datasets collection from UCI Data repository (https://github.com/lpfgarcia/ucipp/blob/master/uci/car-evaluation.arff) having multi attribute values as given in Table-1. It presents the number of attributes and total number data instances have in each dataset. Each attribute of these has a unique set of values which must have to relate to compute the condition probability with these records as given in Table-2.

Table 1. UCI Categorical Datasets

UCI Datasets	Attributes	Data Instances	Clustering Class (k)
Car	6	1728	4
Nursery	8	12960	5
Mushroom	22	8124	2

Table 2. Attribute unique values

Attributes	Unique Attribute values
Buying	v-high, high, med, low
Maintain	v-high, high, med, low
Doors	2,3,4,5-more
Persons	2,4, more
Lug Boot	Small, med, big
Safety	Low, med, high

Evaluating the priorities of clustering is usually trivial and difficult [Brodley and Dy (2004)]. To achieve the highest level of similarity between groups, collectively targeted work is always designed to maintain low relationships. It can be considered as an internal paradigm to evaluate cluster accurateness. However, as per the previous studies suggest that there is no necessitate transforming the excellent results of this function into a higher level of internal reference.

4.2 Measures

The proposed HMARs are compared with the existing approaches with *k*-representatives-Modified [Nguyen et al. (2016)], *k*-representatives [Jiang et al. (2013)] and *k*-mode [Wang et al. (2015)].

Various studies have used purity and NMI to calculate the performance of clustering suggestions in the past. The purity and NMI measures are used to assess the purity and independence of information in a created cluster [Iwata et al. (2018); Tao et al. (2005)]. These approaches are used to assess clustering accuracy based on the relationship between cluster items and the important information of the actual class.

We establish a collection of clusters as |C| with j unit clusters using the clustering technique, and we construct exclusive attribute partitions as |P| with I unit parts across a N number of datasets using an attribute relation matrix.

Purity Measure: It's a metric for determining how accurate data in clusters is in relation to its class. It determines the available data points based on the most desirable features for each unique cluster and computes the amount of data records that are appropriately assigned with the precise match to the class. If the purity result is high, the clustered data created is correct. Eq. (4), as shown below, is used to calculate it

$$Purity(C,P) = \frac{1}{N}\sum_j \max_i \left|C_j \cap P_i\right| \qquad (4)$$

- ***NMI Measure:*** It is widely used to determine the degree of similarity between clusters. NMI has a range of values between 0 and 1. The largest number represents a closer resemblance of values and a lower level of uncertainty, indicating that 1 is the most excellent correlated and perfect cluster,

while 0 shows no mutual link with the maximum level of attribute independence. The Eq. may be used to compute the NMI between clusters (5).

$$NMI(C,P) = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}|C_j \cap P_i| log \frac{N|C_j \cap P_i|}{|C_j||P_i|}}{\sqrt{\sum_{j=1}^{J}|C_j| log \frac{|C_j|}{N} \sum_{j=1}^{J}|P_i| log \frac{|P_i|}{N}}} \quad (5)$$

It was discovered that the number of items, attributes, and values of a variety of data objects in each domain group be at variance from those in other domain groups. When the amount of items and their values in a cluster have a tighter relationship to class, the NMI and Purity act measures are improved and more accurate for their respective category classes.

Precision, recall, entropy, and purity characteristics are used to evaluate outcomes. On the basis of these factors, a comparison of the present technique with the suggested approach is illustrated and explained below.

- ***Precision:*** Precision value is assessed by the recovery of documents at true positive prediction, false positive.

$$\text{Precision} = \frac{\text{No.of relevant documents retrieved}}{\text{Relevant documents in collection}} \quad (6)$$

- ***Recall***: Recall value is assessed by the recovery of documents at true positive prediction, false negative.

$$\text{Recall} = \frac{\text{No.of relevant documents retrieved}}{\text{Relevant Documents in collection}} \quad (7)$$

- ***Entropy:*** The entropy value may be characterised as how much every group contains of data items of a solitary class. The entropy value of a cluster j is decided by utilizing the equation:

$$e_j = -\sum_{i=1}^{L} p_{ij} \log p_{ij} \quad (8)$$

Here L denotes the quality of classes and p_{ij}signifies the probability that an individual from group j has a place with class i.

The overall entropy result is characterised as:

$$e = \sum_{j=1}^{k} \frac{\beta_j}{n} e_j \quad (9)$$

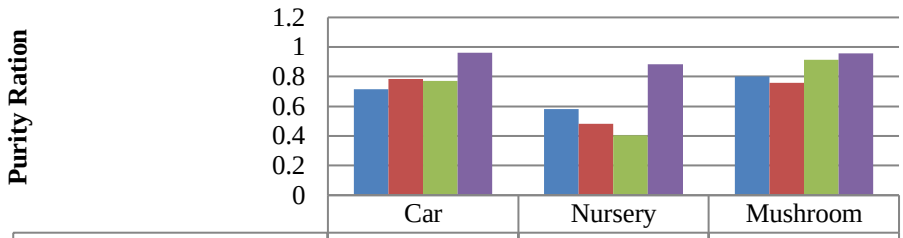
K = the total amount of clusters, n specifies the quantity of input documents in the corpus. Clustering results are considered to be better if the entropy value is low.

5. Result Analysis

In this part, it provides an analysis of the proposed HMARs which is compared with the existing approaches like *k*-representatives-Modified [Nguyen et al. (2016)], *k*-representatives [Jiang et al. (2013)] and *k*-mode [Wang et al. (2015)]. First, the *k*-cluster was equipped with many datasets, and we implemented the HMARs for measuring performance between two dissimilar datasets.

Table 3. Comparison of Purity

Dataset	*k*-mode [Wang et al. (2015)]	*k*-representatives [Jiang et al. (2013)]	*k*-representatives-Modified [Nguyen et al. (2016)]	Proposed HMAR
Car	0.716	0.785	0.771	0.962
Nursery	0.581	0.481	0.405	0.885
Mushroom	0.801	0.759	0.914	0.958



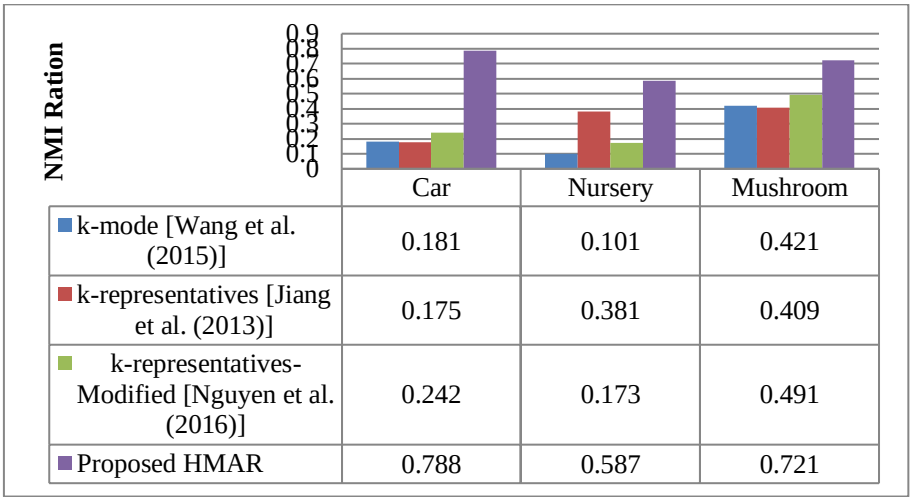
	Car	Nursery	Mushroom
k-mode [Wang et al. (2015)]	0.716	0.581	0.801
k-representatives [Jiang et al. (2013)]	0.785	0.481	0.759
k-representatives-Modified [Nguyen et al. (2016)]	0.771	0.405	0.914
Proposed HMAR	0.962	0.885	0.958

Figure 5. Datasets Purity Result in Comparison

The outcomes in Table-3 show comparative values of the results of measuring purity. In comparison with the *k*-mode and the *k*-representation method, HMARs shows an excellent result for each data set. As shown in Fig. 5, when comparing the purity performance of HMARs with the *k*-mode, it shows ~25%, *k*-representative ~20%, and modified *k*-representative ~12% improvisation.

Table 4. Comparison of NMI Results

Dataset	*k*-mode [Wang et al. (2015)]	*k*-representatives [Jiang et al. (2013)]	*k*-representatives-Modified [Nguyen et al. (2016)]	Proposed HMAR
Car	0.181	0.175	0.242	0.788
Nursery	0.101	0.381	0.173	0.587
Mushroom	0.421	0.409	0.491	0.721



	Car	Nursery	Mushroom
k-mode [Wang et al. (2015)]	0.181	0.101	0.421
k-representatives [Jiang et al. (2013)]	0.175	0.381	0.409
k-representatives-Modified [Nguyen et al. (2016)]	0.242	0.173	0.491
Proposed HMAR	0.788	0.587	0.721

Figure 6. Datasets NMI Result in Comparison

The results in Table-4 and Fig. 6 show a value of comparison that measure of the NMI results. In comparison to HMARs, all the comparing approaches show low NMI values. Accuracy in clustering improvises the NMI values compared to other approaches. It shows an improvisation compared to the *k*-

mode it shows ~20 %, with *k*-representative it shows ~15%, and with modified *k*-representative shows ~20%.

Table 5. Comparison of Proposed HMAR Method Using Different Metrics

Dataset	Precision	Recall	Entropy
CAR	0.880	0.888	0.821
Nursery	0.891	0.885	0.801
Mushroom	0.903	0.900	0.791

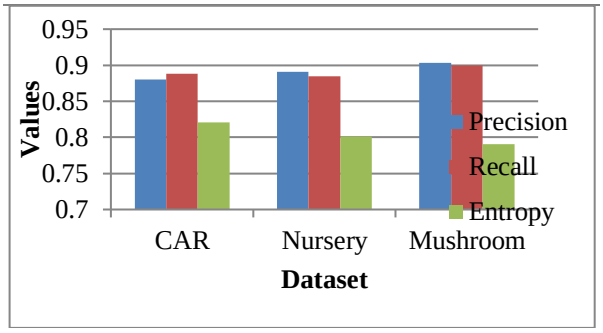


Figure 7. Result Comparison of Hmars Using Various Parameters

In above Table 5 and Fig 7 shows the result comparison of HMARs utilizing various parameters like precision, recall and entropy. It is said that, the proposed HMAR approach provides better results in terms of precision value of 0.903, recall value of 0.900 and entropy value of 0.791 for mushroom dataset.

6. Conclusion And Future Work

In this research, HMARsare suggested that uses co-clustering mechanisms to combine attribute selections with relation analysis. The HMAR is a tool that may be used to find the ARMs (Attribute Relation Matrices) of multiple attribute datasets. It creates ARMs for correct clustering by using clustered data and to create attribute connection models, many attribute values are needed. This work considers the problem of unsupervised feature selections for multiple attribute datasets by employing unique probability attribute connection models to accomplish data reconstructions for clustering. The proposed technique expresses its ability to identify data in actual data spaces while minimizing graphical and data reconstruction errors to retain higher levels of resemblance. HMARs outperform three most current clustering algorithms in experimental tests using multi-valued data sets. From the result analysis, when comparing the purity performance of HMARs with the *k*-mode, it shows ~25%, *k*-representative ~20%, and modified *k*-representative ~12% improvisation. We may enhance the approach to unconditionally assess cluster performance under unsupervised function selection criteria in function.

6.1 Conflict of Interest

The authors of this publication declare there is no conflict of interest.

6.2 Funding Agency

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

[1]. J. Brockmeier, T. Mu, S. Ananiadou, and J. Y. Goulermas, "Self-Tuned Descriptive Document Clustering Using a Predictive Network", IEEE Transactions on Knowledge and Data Engineering, Vol. 30(10), pp.1929 - 1942, 2018.

[2]. W. Hua, Z. Wang, H. Wang, K. Zheng, and X. Zhou, "Understand Short Texts by Harvesting and Analysing Semantic Knowledge", IEEE Transactions on Knowledge And Data Engineering, Vol. 29(3), pp. 499 - 512, 2017.

[3]. H. Jaber, F. Marle, and M. Jankovic, "Improving Collaborative Decision Making in New Product Development Projects Using Clustering Algorithms" IEEE Transactions on Engineering Management, Vol. 62(4), pp. 475 - 483, 2015.
[4]. T. Iwata, T. Hirao, and N. Ueda, "Topic Models for Unsupervised Cluster Matching", IEEE Transactions on Knowledge and Data Engineering, Vol. 30(4), pp.786 - 795, 2018.
[5]. T. -Hien T. Nguyen and V.-N. Huynh, "A k-Means-Like Algorithm for Clustering Categorical Data Using an Information Theoretic-Based Dissimilarity Measure", International Symposium on Foundations of Information and Knowledge Systems, Vol. 9616, pp. 115–130, 2016.
[6]. B. Jiang, J. Pei, Y. Tao, and X. Lin, "Clustering Uncertain Data Based on Probability Distribution Similarity", IEEE Transactions on Knowledge and Data Engineering, Vol. 25(4), pp. 751 - 763, 2013.
[7]. J. Pei, B. Jiang, X. Lin and Y. Yuan, "Probabilistic Skylines on Uncertain Data", International conference on VLDB, pp. 15–26, 2007.
[8]. Y. Tao, R. Cheng, X. Xiao, W.K. Ngai, B. Kao, and S. Prabhakar, "Indexing Multi-Dimensional Uncertain Data with Arbitrary Probability Density Functions", International conference on VLDB, pp. 922–933, 2005.
[9]. Z. Li, Jing Liu, Yi Yang, X. Zhou, and H. Lu, "Clustering-Guided Sparse Structural Learning for Unsupervised Feature Selection", IEEE Transactions on Knowledge and Data Engineering, Vol. 26(9), pp. 2138 - 2150, 2014.
[10]. Y. Yang, H. T. Shen, Z. Ma, Z. Huang and X. Zhou, "L2, 1-norm regularized discriminative feature selection for unsupervised learning", International joint conference on Artificial Intelligence, Vol. 2, pp. 1589–1594, 2011.
[11]. F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selection via joint L2, 1-norms minimization", In Proc. Adv. Neural Information Process System, Vol. 23, pp. 1813-1821, 2010.
[12]. W. Fan, N. Bouguila, and D. Ziou, "Unsupervised hybrid feature extraction selection for high-dimensional non-Gaussian data clustering with variation inference", IEEE Transaction on Knowledge and Data Engineering, Vol. 25(7), pp. 1670-1685, 2013.
[13]. Z. Li, Y. Yang, J. Liu, X. Zhou, and H. Lu, "Unsupervised feature selection using nonnegative spectral analysis", AAAI Conference on Artificial Intelligence, pp. 1026–1032, 2012.
[14]. L. Wolf and A. Shashua, "Feature selection for unsupervised and supervised inference: The emergence of sparsity in a weight-based approach", Journal of Machine Learning Research, vol. 6, pp. 1855-1887, 2005.
[15]. D. Cai, C. Zhang and X. He, "Unsupervised feature selection for multi-cluster data", International Conference of Knowledge Discovery and Data Mining, pp. 333-342, 2010.
[16]. R. Zhao and K. Mao, "Fuzzy Bag-of-Words Model for Document Representation", IEEE Transactions on Fuzzy Systems, Vol. 26(2), pp.794 - 804, 2018.
[17]. X. Pei, C. Chen and W. Gong "Concept Factorization with Adaptive Neighbors for Document Clustering", IEEE Transactions on Neural Networks and Learning Systems Vol. 29 (2), pp.343 - 352, 2018.
[18]. J. Zhu, K. Wang, Y. Wu, Z. Hu, and H. Wang, "Mining User-Aware Rare Sequential Topic Patterns in Document Streams", IEEE Transactions on Knowledge And Data Engineering, Vol. 28(7), pp. 1790 - 1804, 2016.
[19]. X. He, D. Cai and P. Niyogi, "Laplacian score for feature selection", Proceedings of the 18th International Conference on Neural Information Processing Systems, pp. 507–514, 2005.
[20]. J. Wang, J. Wang, J. Song, X.-S. Xu, H. T. Shen, S. Li, "Optimized Cartesian K-Means", IEEE Transactions on Knowledge and Data Engineering, Vol. 27(1), pp. 180 - 192, 2015.
[21]. J. Wu, H. Liu, H. Xiong, J. Cao, and J. Chen, "K-Means-Based Consensus Clustering: A Unified View", IEEE Transactions on Knowledge and Data Engineering, Vol. 27(1), pp. 155 - 169, 2015.
[22]. J. -P. Mei, Y. Wang, L. Chen and C. Miao, "Large scale document categorization with fuzzy clustering", IEEE Transactions on Fuzzy Systems, Vol. 25(5), pp. 1239 - 1251, 2017.

[23]. M. -L. Zhang and Lei Wu, "LIFT: Multi-Label Learning with Label-Specific Features", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 37(1), pp. 107 - 120, 2015.
[24]. Z. Zhao and H. Liu, "Spectral feature selection for supervised and unsupervised learning", In Proc. 24th Int. Conf. Mach. Learn., pp. 1151-1157, 2007.
[25]. X. Li and Y. Pang, "Deterministic column-based matrix decomposition", IEEE Transactions on Knowledge and Data Engineering, Vol. 22(1), pp. 145-149, 2010.
[26]. S. Boutemedjet, D. Ziou, and N. Bouguila, "Unsupervised feature selection for the accurate recommendation of high-dimensional image data", Proceedings in Advances in Neural Information Processing Systems, Vol. 20, 2007.
[27]. E. Brodley and J. G. Dy, "Feature selection for unsupervised learning", International Journal Machine Learning Res., Vol. 5, pp. 845-889, 2004.
[28]. A. M. Almalawi, A. Fahad, Z. T., Muhammad A. Cheema, and I. Khalil "kNNVWC: An efficient k-nearest neighbors approach based on various-widths clustering", IEEE Transactions on Knowledge and Data Engineering, Vol. 28(1), pp. 68 - 81, 2016.
[29]. D. Ienco, R. G. Pensa and R. Meo, "From context to distance: learning dissimilarity for categorical data clustering", ACM Trans. Knowledge Discovery Data, Vol. 6(1), pp.1-25, 2012.
[30]. O. M. San, V. N. Huynh, and Y. Nakamori, "An alternative extension of the k-means algorithm for clustering categorical data", International Journal Application Mathematics, Computation Science, Vol. 14, pp. 241- 247, 2004.
[31]. L. Chen, Q. Jiang, and S. Wang, "Model-Based Method for Projective Clustering", IEEE Transactions on Knowledge and Data Engineering, Vol. 24(7), pp. 1291 - 1305, 2012.
[32]. Natthakan I.-On, T. Boongeon, S. Garrett and C. Price, "A Link-Based Cluster Ensemble Approach for Categorical Data Clustering", IEEE Transactions on Knowledge And Data Engineering, Vol. 24(3), pp. 413 - 425, 2012.
[33]. J. Tang, X. Hu, H. Gao and H. Liu, "Discriminate analysis for unsupervised feature selection", International Conference on Data Mining, pp. 938-946, 2014.
[34]. H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering", IEEE Transactions on Knowledge and Data Engineering, Vol. 17(4), pp. 491-502, 2005.