Associative Rule based Fuzzy Clustering Performance and their Evaluation on Ecommerce Historical data

Anima P[1], Dr A.S. Aneeshkumar[2]

[1]Research Scholar, Department of Computer Science, AJK College of Arts and Science, Coimbatore, Tamilnadu - 641105, India

[2]Research Supervisor & Head, Research Department of Computer Science & Applications,
AJK College of Arts and Science, Coimbatore, Tamilnadu - 641105, India

Abstract

The clustering technology that enables the user to manage massive amounts of data effectively, the purpose of clustering is to convert data from any source into a more compact form that correctly captures the original material.

The user should be able to manage and make better use of the original volume of data, because it would be ineffective if the compact form of the data did not precisely reflect the original data, clustering accuracy is crucial. The accuracy of a well-known fuzzy clustering technique is one of our primary contributions.

It is difficult to examine and implement association rule mining because of the sheer number of rules that are produced from the dataset. For handling association rules, a novel hybrid method called ARFC—Association Rules Fuzzy Clustering—is put forth.

Keywords: Fuzzy clustering analysis; Association rule; Data mining; historical data base.

INTRODUCTION

Clustering is a crucial and widely used method for automatically extracting knowledge from massive volumes of data. Exploring the distribution of items in a data collection is its task. A data collection is often divided into groups (clusters) using the clustering approach, which aims to make the data items allocated to the same cluster as similar as feasible, and the data items given to other clusters as dissimilar as possible.

Data clustering is utilised in a variety of industries, including bioinformatics for the study of microarray data and database marketing for the purpose of consumer segmentation. The clustering approaches may be applied to other activities in the same field. For instance, clustering techniques are frequently employed for text summarization in text mining.

Huge volumes of data may be analysed using association rules to find correlations and possible linkages between different objects or features. These guidelines can be useful in revealing previously undiscovered linkages and producing findings that can serve as the foundation for predictions and decisions. They have shown to be quite helpful. One of the most well-known subfields of data mining research is the use and development of association rules [1].

The purpose of this study is to create and construct an inference mechanism for association rule mining, in order to uncover abstract information from enormous number of frequent patterns. In this study, we suggested brand-new algorithm, ARFC, to accomplish the desired result.

OBJECTIVES OF THIS RESEARCH

- To select an appropriate association rule mining strategy for identifying frequent patterns
- To create a framework of inference mechanisms for association rule mining.
- To apply this mechanism as an inference to find information based on the inference from frequent patterns.
- To do an analysis to satisfy knowledge of abstract inference.

LITERATURE REVIEW

Subdividing data items into subsets is the procedure that used in clustering analysis. Each subset is a cluster, and items inside clusters have a lot in common while also diverging quite a bit from one another. Clustering techniques are widely used in a variety of fields, including engineering, biology, biometric data, sentiment analysis, and others [2].

Data sequences have characteristics including nonlinearity, high dimension, complexity, and redundancy. Clustering is a key component of the data mining method, which is used to analyse data and find probable rules. Data clustering is directly impacted by three factors: accurate clustering number acquisition, an exceptional and effective clustering method, and a distance function that measures how similar the data points are to one another [3]. Gupta and Srivastava proposed a document-based K-means technique that chooses K and refined clusters automatically to address the problem of cluster number selection. [4].

Onan and Toolu [5]. provides an unbalanced learning method based on consensus clustering and under sampling. This plan uses five supervised learning techniques. An enhanced framework for integrating clustering that incorporates clustering techniques. More than three Embedding techniques, weighing and clustering algorithms are used in their study to cluster documents.

Ho et al summary's of fuzziness theory-based data mining research, an explanation of recent research findings on fuzzy clustering techniques, and a research outlook are provided [6]. The fuzzy C-mean (FCM) approach for fuzzy clustering is the most typical one. Fuzzy theory is applied to the data through the fuzzy C-mean technique. This approach is a flexible partition that can gather a lot of clustering data and more precisely reflect the sample distribution in real life.

An enhanced fuzzy C-means approach used a combination of the genetic algorithm, clustering swarm optimization algorithm was developed by Rohidin. [7]. The improved method lessened the algorithm's reliance on the original centre point and somewhat improved the performance of the clustering process.

Sanz [8] recommended a system that used utilised the midpoints of two closest sample points as the cluster centre points in their novel method for selecting cluster centers. Li, Yafang's [9] clustering approach efficiently grouped enormous data by dividing it into smaller groups and resolving the issue of overlapping clustering centers.

An efficient clustering approach has been developed by Bulut et al. [10] for microarray gene expression data. The problem that occurs in the field of bioinformatics is resolved by the algorithm using the feature selection approach. The fuzzy algorithm and the merging are used to improve the quality of the ant-based clustering approach. In Rammal et al [11], the experimental findings employing the L1 or Manhattan distance were superior to the Euclidean and Cheyne distances, according to research on infrared spectrum clustering. Izakian et al. [12] successfully implemented the average value approach to the Fuzzy clustering algorithm.

An updated method was proposed by Abhirami et al. [14], and experimental findings revealed that the modified fuzzy cluster algorithm was more accurate. According to Balaji Padmanabhan et al. [15], association rule mining are produced from the datasets, making it challenging for users to value the use of the rules, to resolved using constraint-based data mining, post-pruning rules, grouping rules, and unpredicted patterns [16].

Principle of ARFC algorithm

In our method, the entropy values of the data points (historical data) are calculated, and the data point with the lowest entropy value is selected as the cluster centre [17]. In this iterative method, the data points are sorted according to a threshold degree of similarity.

Below is an explanation of the ARFC algorithm's basic concept. Assume there are dN data points in M-dimensional [mT] hyperspace, each of which is represented by a collection of M values at the level of Li I = (1, 2, 3,..., N) (i.e., Li1, Li2, Li3, . . . , LiM). Thus, an N M matrix may be used to represent the data collection.

$$D_{ij} = \sqrt{\sum_{k=1}^{M} (X_{ik} - X_{jk})^2}.$$

For each of Dij, the maximum N2 distance values are N C2 distances. There are N diagonal values, all of which are equal to zero (where i = j). The following formula may be used to calculate how similar any two points (i and j) are to one another.

$$S_{ij} = e^{-\alpha \ D_{ij}},$$

Where, the constant represents a number. It should be remembered that any two points have similarity value between them that ranges from 0.0 to 1.0.

$$\overline{D} = \frac{1}{{}^{N}C_2} \sum_{i=1}^{N} \sum_{j>i}^{N} D_{ij}.$$

α can be determined as follows:

$$\alpha = -\frac{\ln 0.5}{\overline{D}}.$$

Then, each data point's entropy (Eni) in relation to the other data points is calculated as follows:

$$E_i = -\sum_{j \in X}^{j \neq i} (S_{ij} \log_2 S_{ij}) + (1 - S_{ij}) \log_2 (1 - S_{ij}))$$

During clustering, the data point with the lowest entropy value is selected as the cluster centre. An explanation of the clustering algorithm is provided below.

The clustering algorithm entails the following actions:

1. For each Li located in [Thp] hyperspace, determine Ei =i1, i2, i3,..., N).

2. A cluster centre of Li,M is chosen after determining the minimal Ei.

3. Put Li, Min and the data points in a cluster that are more similar to Li, Min than the similarity criteria (), then remove those data points from [Thp].

4. Verify the emptiness of [Thp] hyperspace. If so, stop the programme; otherwise, move on to Step 2.

This way of determining Eni allows even a data point furthest from the other data points to be selected as the cluster centre. In order to prevent this, another option called (in%), which is just a threshold used to identify a cluster as a legitimate one, has been added. If there are more than or equal to N 100 data points in a cluster, we consider the cluster to be valid. If not, these data points will be treated as outliers. The ARFC algorithm has been extended as mentioned.

Association Rule Algorithm Principles

An association between two sets of items are said to exist, when a transaction containing one set of items is likely to also contain in other set. For instance, "67% of all consumers who buy milk also buy curd" may be found through the study of past e-commerce data.

A rule that is latent in databases and may be extracted from them is an association [18]. It makes possible to infer one attribute set from another. Association rules may be viewed as a technique for obtaining meaningful information by extracting patterns from very big databases. The primary objective of association-rule mining is to identify rules that have the minimal needed support and confidence level.

The selection of all objects (itemsets) with a support level greater than the bare minimum required support level is the first step in association-rule mining. The second step is the finding of interesting rules from these itemsets[19]. Datasets that may not contain class labels can nonetheless extract associations between attributes using association rules. Techniques for extracting association rules include most frequent itemset to identify connections between items in transaction data.

If the fractions of the transactions in D contain (XY), the rule XY has support S in D. The goal of mining association rules is to provide all association rules with a minimal level of support (min-sup) and confidence (min-conf) set by the user.

Support (XY) = frequency (X) / total records in the database

The degree of association between X and Y in the database is indicated by the confidence (C) of a rule. Confidence (X)=frequent (X)=frequent (Y)/frequent (X)

The strength of the regulations is also measured by confidence. Mining is the process of identifying all rules that meet the user-specified threshold support and confidence [19].

Fuzzification

A real scalar value is transformed into a fuzzy value through the process of fuzzyfication. The various varieties of fuzzifiers are used to achieve this. The first stage of fuzzy inferencing is fuzzyification. Crisp inputs are converted into fuzzy inputs in this domain transformation. [20]. Each crisp input that the fuzzification inference unit needs to process has a unique set or group of membership functions to which it is changed. This membership group operates inside a discourse universe that encompasses all pertinent values that the fresh input may have. The organisation of membership functions inside a realm of discourse for a crisp input is depicted in Figure 1.

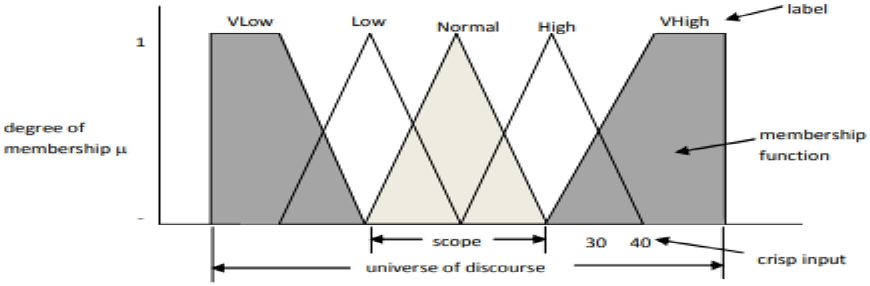
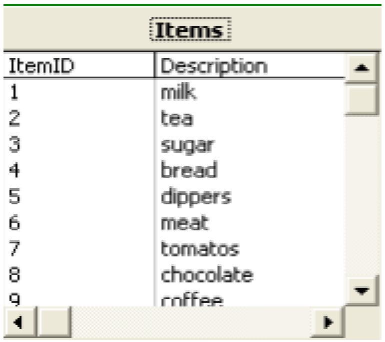


Figure 1: Structure of Membership function for crisp input

Min Conf is the minimal confidence threshold, whereas Min Supp is the minimum support level. The user must provide both of them to execute association Rule.

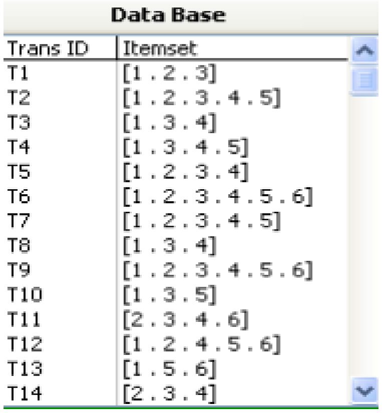


Items

ItemID	Description
1	milk
2	tea
3	sugar
4	bread
5	dippers
6	meat
7	tomatos
8	chocolate
9	coffee

Figure 2: Preferred purchased items table

Figure 3 shows the transactions between products, i.e., the items a customer purchases in combination. For example, transaction 1 indicates that a client buys milk, tea, and sugar together, whereas transaction 2 indicates that a consumer buys milk, tea, sugar, bread, and dippers together.

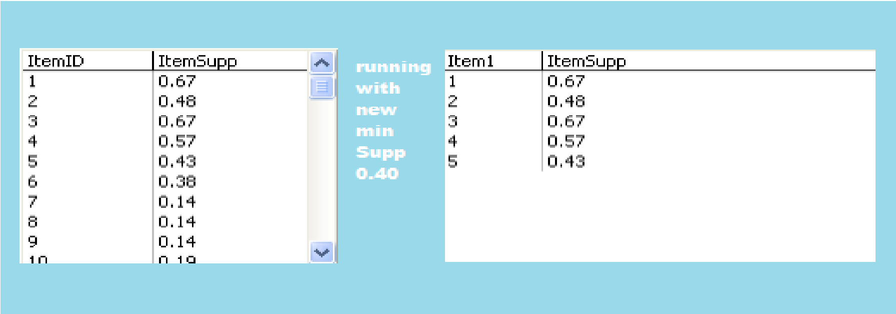


Data Base

Trans ID	Itemset
T1	[1 . 2 . 3]
T2	[1 . 2 . 3 . 4 . 5]
T3	[1 . 3 . 4]
T4	[1 . 3 . 4 . 5]
T5	[1 . 2 . 3 . 4]
T6	[1 . 2 . 3 . 4 . 5 . 6]
T7	[1 . 2 . 3 . 4 . 5]
T8	[1 . 3 . 4]
T9	[1 . 2 . 3 . 4 . 5 . 6]
T10	[1 . 3 . 5]
T11	[2 . 3 . 4 . 6]
T12	[1 . 2 . 4 . 5 . 6]
T13	[1 . 5 . 6]
T14	[2 . 3 . 4]

Figure 3: transactions between products, customer purchases in combination

As illustrated in figure 3, the user selects ARFC and types a user threshold number into the (User Thr.). The user selects ARFC and enters a user threshold number as shown in Figure (3). (User Thr.). If the user enters 0.3 in the (User Thr.) box, the same number for the (Min Sup) and (Min Conf) boxes, and clicks on the (Find Ass Rules). Figure 4 shows how C1 and L1 are constructed using the same methods.



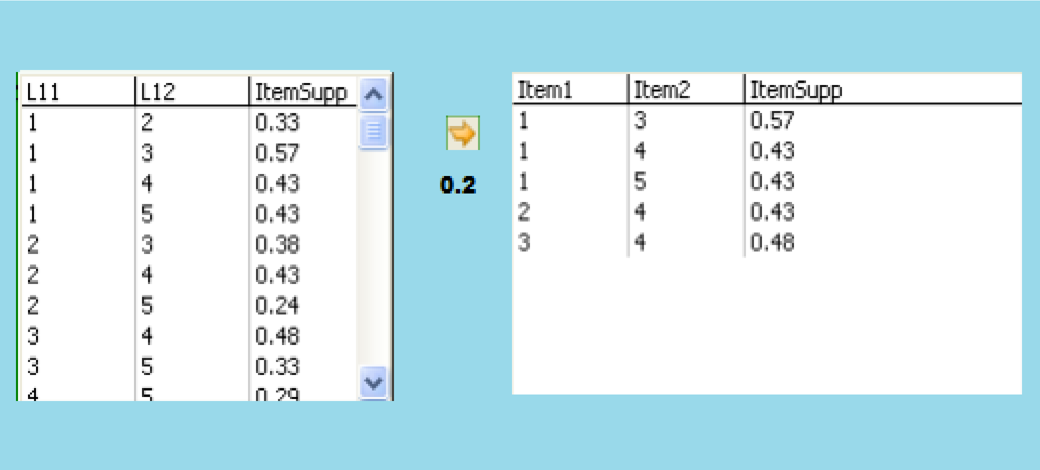
ItemID	ItemSupp
1	0.67
2	0.48
3	0.67
4	0.57
5	0.43
6	0.38
7	0.14
8	0.14
9	0.14
10	0.19

running with new min Supp 0.40

Item1	ItemSupp
1	0.67
2	0.48
3	0.67
4	0.57
5	0.43

Figure 4: Item running with new minimum support of 0.40

Using 0.40, a new minimum support is determined. The construction of C2 and L2 is dependent on the new minimum support of 0.40; all items with Item Support greater than or equal to 0.40 survive, while all other items are trimmed, as shown in figure (5).

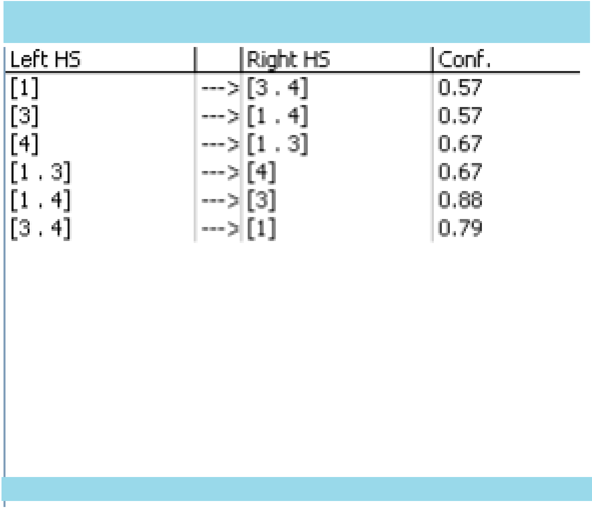


L11	L12	ItemSupp
1	2	0.33
1	3	0.57
1	4	0.43
1	5	0.43
2	3	0.38
2	4	0.43
2	5	0.24
3	4	0.48
3	5	0.33
4	5	0.29

Item1	Item2	ItemSupp
1	3	0.57
1	4	0.43
1	5	0.43
2	4	0.43
3	4	0.48

Figure 5: Item running with new minimum support of 0.2

0.2 is the new minimal support. As shown in figure, all things with ItemSupp greater than or equal to 0.2 are kept, while all other items are removed. C3 is constructed in the same way, but L3 must meet the increased minimum support requirement of 0.2.



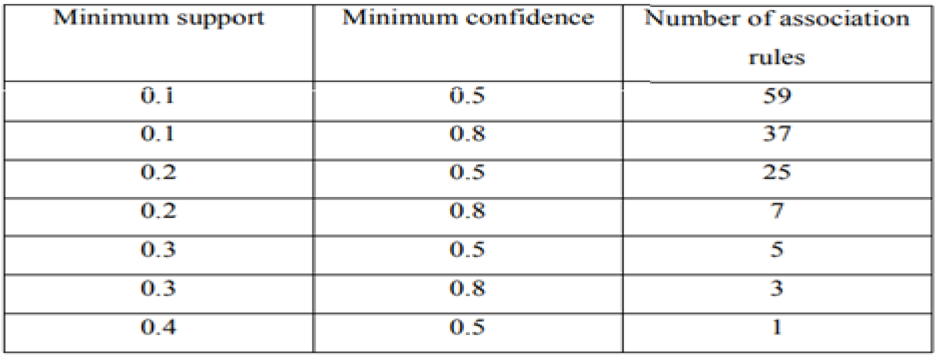
Left HS		Right HS	Conf.
[1]	--->	[3 . 4]	0.57
[3]	--->	[1 . 4]	0.57
[4]	--->	[1 . 3]	0.67
[1 . 3]	--->	[4]	0.67
[1 . 4]	--->	[3]	0.88
[3 . 4]	--->	[1]	0.79

Figure 6: Item running with minimum confidence

Only the rules that have minimal confidence above 0.85 remain after all other rules are pruned, as shown in figure 8. The new minimum confidence is calculated using association rules generated, and the obtained rules are only those that have confidence above the (New Min Conf).

IMPLEMENTATION RESULTS

Table -1 ARF Cresults - confidence increase



Minimum support	Minimum confidence	Number of association rules
0.1	0.5	59
0.1	0.8	37
0.2	0.5	25
0.2	0.8	7
0.3	0.5	5
0.3	0.8	3
0.4	0.5	1

The results of the implementation demonstrate that the number of association rules decreases as support and confidence rise.

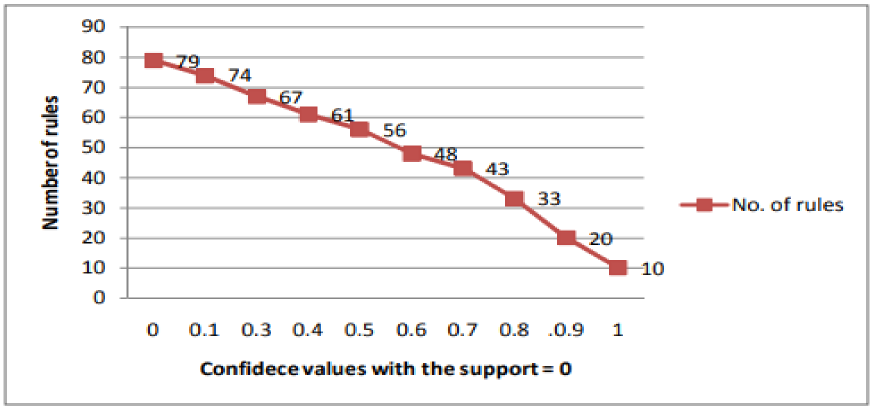


Figure 7. The number of rules generated against the confidence

The implementation results demonstrate that the number of association rules decreases when support, confidence, and user thresholds rise.

Application of Association Rules Fuzzy Clustering algorithm (ARFC)

Python programming is sometimes said to as "high-level," which means that the programmer need not worry about things like direct memory management. The built-in features of the language handle everything, allowing the programmer to focus just on the elements of the issue at hand. But in programming, everything is relative, just like in many other aspects of life. Programming in "pure" Python is therefore seen as being excessively "low level" for some tasks, such as data processing. This is because principles from linear algebra are heavily used in all scientific domains that deal with data.

The main justification for this is that the matrix, which is the fundamental structure established in linear algebra, offers a perfect framework for structuring and modifying the data under study via matrix operations. Data analysis, which includes cluster analysis as a discipline, is not an exception to this norm. The application of linear algebra increases the demand for a programmatic implementation of the matrix structure. However, the addition of a matrix structure to a program significantly affects the entire method of handling the data. The software now manipulates data as a group of components that are stored in matrices rather than as individual items that are handled separately.

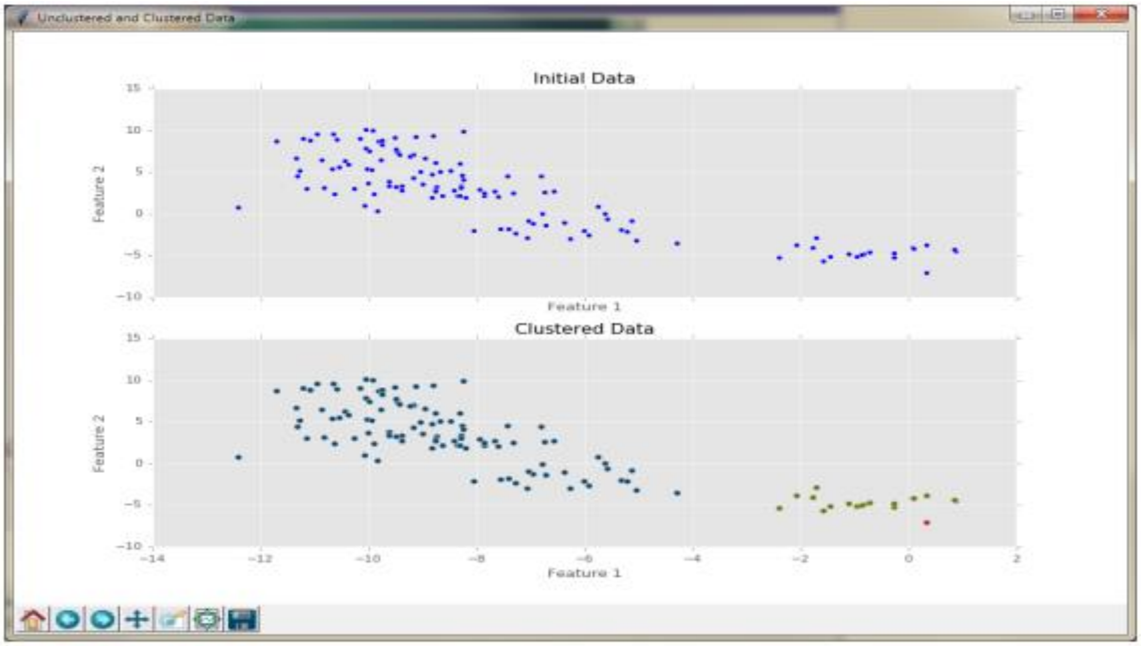


Figure- 8 Cluster analysis

Nodes of colours other than blue can be seen all around a single cluster, indicating that they are members of various clusters. One solution would be to do a merging operation between such clusters with a limited number of nodes, such as 1–5, following the ARFC execution and assign them to the nearby bigger cluster. Prior to modifying the parameters of the method, we sequentially perform the relative criteria indices on the dataset. We have decided to implement indices designed for this sort of clustering as the ARFC method is a hard clustering algorithm. However, unlike the k-means algorithm, which varies the number of clusters.

The clustering in every one of the evaluated k and f combinations that maximize the indices has been verified to be the same and the best that could be achieved. You may view it below:

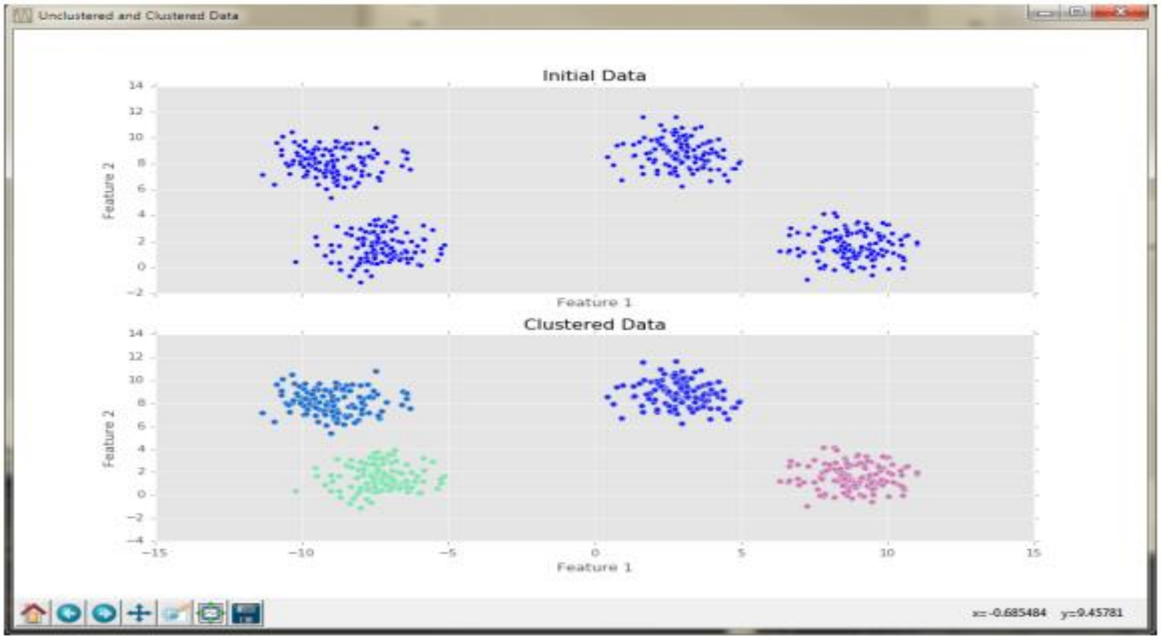


Figure 9 - Execution of ecommerce data

The best clustering that ARFC is seen in figure 9. Because there are certain nodes that operate as a "bridge" between the two clusters on the right, making it difficult for the algorithm to separate them without disrupting the remainder of the partition, no other combination of parameters can lead the indices at better values.

Although some of the vectors are closer to the centroid of the cluster with blue colour, we can see from the first scatter plot above that some of the vectors belong to the cluster with green colour. As previously said, this occurs because these green coloured vectors were initially allocated to the green cluster, but towards the conclusion of the algorithm's execution, the centroid of the pink cluster very near approached them.

The best clustering that ARFC is capable of is seen in figure 9. Because there are certain nodes that operate as a "bridge" between the two clusters on the right, making it difficult for the algorithm to separate them without disrupting the remainder of the partition, no other combination of parameters can lead the indices at better values.

CONCLUSIONS

The following is a list of the conclusions that might be derived from this work:

The number of association rules created with fuzzy is lower than the number of association rules with the novalapproach. When a fuzzy function is used, the transaction database is filtered in accordance with a predetermined threshold (this threshold specifies the maximum number of items in the transaction), which helps to save space and speed up processing.

- If Min Sup is small in the ARFC algorithm, a significant number of association rules are created.
- If Min Sup and User Thr. are small in Function1, a lot of association rules are created.
- If Max Sup and User Thr. are high, there are many association rules created in Function 2.

On historical datasets, we investigated, put our clustering methods to the test, and made sure they met all the necessary cluster validity requirements. Finally, we used the technique we had suggested to the task of segmenting the data base, and we produced a successful outcome.

The area of unsupervised learning known as cluster analysis contains procedures that categorise data based on proximity. All processes used to assess a clustering algorithm's output are collectively referred to as cluster validity. We looked at, put into practice, and evaluated ARFC clustering algorithms using historical datasets with all the necessary cluster validity requirements. Last but not least, we applied proposed algorithms to the problem of segmenting purchased goods and put forth a useful method for compacting purchased items on clustered data.

REFERENCES

[1] Tsay, Yuh-Jiuan, and Jiunn-Yann Chiang. “CBAR: An Efficient Method for Mining Association Rules.” Knowledge-Based Systems, vol. 18, no. 2–3, Apr. 2005, pp. 99–105. DOI.org (Crossref), https://doi.org/10.1016/j.knosys.2004.04.010.

[2] Onan, Aytug. “Two-Stage Topic Extraction Model for Bibliometric Data Analysis Based on Word Embeddings and Clustering.” IEEE Access, vol. 7, 2019, pp. 145614–33. DOI.org (Crossref), https://doi.org/10.1109/ACCESS.2019.2945911.

[3] Davidson, James E. H., et al. “Time Series Modeling of Paleoclimate Data: PALEOCLIMATE DATA.” Environmetrics, vol. 27, no. 1, Feb. 2016, pp. 55–65. DOI.org (Crossref), https://doi.org/10.1002/env.2373.

[4] Dogan, Onur, et al. “Fuzzy Association Rule Mining Approach to Identify E-Commerce Product Association Considering Sales Amount.” Complex & Intelligent Systems, vol. 8, no. 2, Apr. 2022, pp. 1551–60. DOI.org (Crossref), https://doi.org/10.1007/s40747-021-00607-3.

[5] Onan, Aytuğ, and Mansur Alp Toçoğlu. “Weighted Word Embeddings and Clustering-based Identification of Question Topics in MOOC Discussion Forum Posts.” Computer Applications in Engineering Education, vol. 29, no. 4, July 2021, pp. 675–89. DOI.org (Crossref), https://doi.org/10.1002/cae.22252.

[6] Ho, G. T. S., et al. “Using a Fuzzy Association Rule Mining Approach to Identify the Financial Data Association.” Expert Systems with Applications, vol. 39, no. 10, Aug. 2012, pp. 9054–63. DOI.org (Crossref), https://doi.org/10.1016/j.eswa.2012.02.047.

[7] Rohidin, Dede, et al. "Association Rules of Fuzzy Soft Set Based Classification for Text Classification Problem." Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 3, Mar. 2022, pp. 801–12. DOI.org (Crossref), https://doi.org/10.1016/j.jksuci.2020.03.014.

[8] Sanz, J., et al. "A Fuzzy Association Rule-Based Classifier for Imbalanced Classification Problems." Information Sciences, vol. 577, Oct. 2021, pp. 265–79. DOI.org (Crossref), https://doi.org/10.1016/j.ins.2021.07.019..

[9] Li, Yafang, and Yisong Li. "E-Commerce Order Batching Algorithm Based on Association Rule Mining in the Era of Big Data." 2018 Chinese Control And Decision Conference (CCDC), IEEE, 2018, pp. 1934–39. DOI.org (Crossref), https://doi.org/10.1109/CCDC.2018.8407443.

[10] Bulut, Hasan, et al. "An Improved Ant-Based Algorithm Based on Heaps Merging and Fuzzy c-Means for Clustering Cancer Gene Expression Data." Sādhanā, vol. 45, no. 1, Dec. 2020, p. 160. DOI.org (Crossref), https://doi.org/10.1007/s12046-020-01399-x.

[11] Rammal, Abbas, et al. "Optimal Preprocessing and FCM Clustering of MIR, NIR and Combined MIR-NIR Spectra for Classification of Maize Roots." The Third International Conference on E-Technologies and Networks for Development (ICeND2014), IEEE, 2014, pp. 110–15. DOI.org (Crossref), https://doi.org/10.1109/ICeND.2014.6991363.

[12] Izakian, Hesam, et al. "Fuzzy Clustering of Time Series Data Using Dynamic Time Warping Distance." Engineering Applications of Artificial Intelligence, vol. 39, Mar. 2015, pp. 235–44. DOI.org (Crossref), https://doi.org/10.1016/j.engappai.2014.12.015.

[13] Abhirami, K. "Web Usage Mining Using Fuzzy Association Rule." 2016 International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS), IEEE, 2016, pp. 1–4. DOI.org (Crossref), https://doi.org/10.1109/ICETETS.2016.7603022..

[14] Padmanabhan, Balaji, and Alexander Tuzhilin. "Unexpectedness as a Measure of Interestingness in Knowledge Discovery." Decision Support Systems, vol. 27, no. 3, Dec. 1999, pp. 303–18. DOI.org (Crossref), https://doi.org/10.1016/S0167-9236(99)00053-6.

[15] Yih-Jen Horng, et al. "A New Method for Fuzzy Information Retrieval Based on Fuzzy Hierarchical Clustering and Fuzzy Inference Techniques." IEEE Transactions on Fuzzy Systems, vol. 13, no. 2, Apr. 2005, pp. 216–28. DOI.org (Crossref), https://doi.org/10.1109/TFUZZ.2004.840134..

[16] Łuczak, Maciej. "Hierarchical Clustering of Time Series Data with Parametric Derivative Dynamic Time Warping." Expert Systems with Applications, vol. 62, Nov. 2016, pp. 116–30. DOI.org (Crossref), https://doi.org/10.1016/j.eswa.2016.06.012.

[17] Huang, Xiaohui, et al. "Time Series k -Means: A New k -Means Type Smooth Subspace Clustering for Time Series Data." Information Sciences, vol. 367–368, Nov. 2016, pp. 1–13. DOI.org (Crossref), https://doi.org/10.1016/j.ins.2016.05.040.

[18] Ferreira, Leonardo N., and Liang Zhao. "Time Series Clustering via Community Detection in Networks." Information Sciences, vol. 326, Jan. 2016, pp. 227–42. DOI.org (Crossref), https://doi.org/10.1016/j.ins.2015.07.046.

[19] Brin, Sergey, et al. "Dynamic Itemset Counting and Implication Rules for Market Basket Data." Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data - SIGMOD '97, ACM Press, 1997, pp. 255–64. DOI.org (Crossref), https://doi.org/10.1145/253260.253325.

[20] Castro-Schez, Jose Jesus, et al. "A Highly Adaptive Recommender System Based on Fuzzy Logic for B2C E-Commerce Portals." Expert Systems with Applications, vol. 38, no. 3, Mar. 2011, pp. 2441–54. DOI.org (Crossref), https://doi.org/10.1016/j.eswa.2010.08.033.