Nat. Volatiles & Essent. Oils, 2021; 8(5): 8304 - 8311

NVEO
Natural Volatiles & Essential Oils

Unsupervised Modified Clustering Technique Based on Fuzzy Set Theory to Categorize the Real-Life Data

R Devi
Department of Mathematics
Pachaiyappa's College, Chennai, India

ABSTRACT
The fuzzy C-means clustering algorithm is appropriate for segmenting datasets and is commonly utilized in real-life settings. The growth of huge data has posed numerous obstacles for clustering approaches. Due to their high density and execution time, traditional clustering approaches cannot be applied for such vast amount of real-world data. To address these challenges, we offer an unsupervised clustering method for automatically categorizing large-scale datasets without requiring labels, with a focus on the real-life dataset. Experiments on real-world datasets show that our suggested unsupervised technique performs well and has high precision. The outcomes reveal that the proposed approach effectively segregate unstructured real-world database into distinct clusters.
Keywords: Clustering, Complex structure, Fuzzy C-Means, Uncertainty

1. Introduction

Clustering analysis is the act of discovering data objects and grouping things altogether based on its characteristics. A cluster is a collection of things that are very similar to other clusters. As a result, clustering [3] is classified as an unsupervised process because it does not have a previous group ID. The center can have attributes of the same dimensions as the object in the data. Alternatively, it can be a symmetric high-level element such as a linear or non-linear subspace or function. Although clusters can be thought of as segments of a larger data set, one categorization method could be based on the fuzzy or crispness of the subsets. The hard clustering approach [17] is based on set theory and requires data regardless of whether it belongs to a cluster or not. Most clustering approaches assume that the clusters are well defined and that each pattern can belong to only one cluster [1]. This option can overlook the natural ability of data to be distributed in different clusters. Fuzzy clustering [11] can be used to overcome the weaknesses in this case by using fuzzy logic. Soft clustering algorithm is more natural than crisp clustering because elements at the boundaries of many clusters are not forced to belong entirely to any of the clusters. The clustering approach [19] has been used in a variety of real-world situations. Because there is no clear demarcation between groups in many real-world scenarios, fuzzy clustering is better suited to the data. Fuzzy set-based clustering [18] is a great way to extract features from data elements that have a local structure. Fuzzy clustering techniques are used to show the local structure of a dataset by predicting membership degree of every data element in the cluster. In fuzzy set theory [7, 13], membership values are often associated with a degree of membership. These membership-level assignments and their subsequent use for allocating data components to clusters are known as fuzzy clustering [5]. Instead of uniquely assigning objects to the cluster, fuzzy clustering uses a membership level of 0 to 1. Such membership can be used to undertake soft data analysis that considers non-linear data structures. It is, however, merely bounded at local minima and is sensitive to changes in the environment [4,9]. Krishnapuram and Keller [6] proposed the PCM technique, which simplifies the probabilistic requirement and permits a possibilistic view of the membership equation as a degree of typicality. The PCM results, are extremely dependent on the initialization and frequently depreciate owing to the overlap clustering problem. Pal et al [8] introduced a fuzzy induced possibilistic c-means clustering approach that addresses the flaws in both FCM and PCM algorithms. Even if the FPCM is the combination of FCM [14] or PCM, the typicality value becomes very low as the size of the dataset rises. As a result, this task seeks to provide an effective clustering technique for evaluating dataset by merging membership values, typicality, and distances guided by the Cauchy kernel. This work is structured as follows: The Preliminaries are given in Section

2 of this study. The proposed technique is given in section 3. Section 4 explains the experimental results of artificial and benchmark data. Finally, Section 5 brings the conclusion.

2. Preliminaries

2.1 K-Means Clustering

K-means clustering is a very well-known unsupervised machine learning algorithm. It is used to solve many problems of unsupervised machine learning. The K-means clustering algorithm attempts to group similar elements in the form of a cluster. The number of groups is represented by K. K-means clustering aims to reduce distances inside a cluster while increasing distances between clusters. It is an iterative procedure to use K-means. The approach is based on the optimization method. It functions by doing the below stages after fix the number of clusters:

1. Select prototypes at arbitrary for each group.

2. Find the detachment between each data elements and the prototypes.

3. Allocate data elements to the group that would be nearest to them.

4. Locate individual group's new prototypes.

5. Reiterate the steps 2, 3, and 4 till all data elements have converged and the center of the cluster has stopped moving.

2.2 Fuzzy Set

A fuzzy set F of a set U can be defined as a set of well-ordered pairs $\{(x, \chi_A(x)): x \in U\}$,each with the first data from U and the next data from the interval [0, 1] by precisely 1 ordered pair exist for each element of U. This defines a mapping, μ_F among objects of the set U and degrees in the interval [0, 1]:

$$\mu_F: U \rightarrow [0, 1].$$

The degree zero is used to represent exact non-membership, the degree one is used to represent exact membership and values among the interval are used to denote transitional grades of membership.

2.3 Kernel Distance

Kernel-induced distance is an effective way to extract information from high-dimensional data by transforming elements from small-dimensional to high-dimensional space [10]. For all mathematical methods that can be specified in a dot product relationship, the mapping provides a linear to non-linear connection. In functional space, the kernel is described as an inner product. The map transforms the n-dimensional data into the inner product of the feature space, and kernel resembles to the inner-product of the feature-space. In this study, to calculate the dot product, we describe the inner product space of the kernel as follows:
$A(u, v) = \langle \pi(u), \pi(v) \rangle$ The above procedure is used for calculating the value of the inner product of the feature space. The kernel induced distance is defined as $\| \pi(u)\text{-}\pi(v) \|^2 = \langle \pi(u)\text{-}\pi(v), \pi(u)\text{-}\pi(v) \rangle =$

$A(u,u)+A(v,v)-2A(u,v)$. Therefore, the kernel induced distance is as $\left\| \pi(u)-\pi(v) \right\|^2 = 2-2A(u,v)$, since $A(u,u)=1$& $A(v,v)=1$. $\left\| \pi(u)-\pi(v) \right\|^2 = 2\left(1-A(u,v)\right)$

3 Modified Possibilistic Kernelized Fuzzy Clustering with Weighted term (MPKFCW)

To deal the outlier problem and heavy noise in segmenting real world dataset [12], the Modified kernelized Fuzzy Clustering algorithm is given in this section. The proposed technique is developed by incorporating the Weighted term. The Weighted term is effectively regularized the clustering process and is worked as fuzzifier of the system.
The new objective function is formulated as

$$O(m,p)=2\sum_{p=1}^{n}\sum_{j=1}^{c}\left(m_{jp}+p_{jp}^{\alpha}\right)\left(1-C_w(z_p,v_j)\right)+\frac{2G}{n}\sum_{j=1}^{n}\sum_{p=1}^{c}m_{ik}\,log\left(\frac{m_{jp}}{C_j}\right) \quad (1)$$

where where $C_w(z_p,v_j)=1-\left(1+\gamma\left\|z_p-v_j\right\|^2\right)^{-1}$, and γ denotes the regularization parameter. *G* is the Geometric mean value of all detachment between the data and prototype and $C_i$is the probabilistic weight of the i[th] cluster. The following equality constraints are used to optimizes this problem

$$\sum_{j=1}^{c}m_{jp}=1\ ,\ \sum_{p=1}^{n}p_{jp}=1\ \&\ \sum_{j=1}^{c}C_j=1 \quad (2)$$

3.1 Membership Function

To obtain the effective membership function to measure the degree of similarity between the data and prototype, the proposed model is minimized subject to the membership constraint.
The general equation for updating membership function is attained as

$$m_{jp}=\frac{C_j\,exp\left[C_w(z_p,v_j)\frac{G}{n}\right]}{\sum_{l=1}^{C}C_l\,exp\left[C_w(z_p,v_l)\frac{G}{n}\right]} \quad (3)$$

3.2 Prototype Equation

Optimizing the *MPKFCW* objective function, the center v_j is evaluated. The cluster center is given by

$$v_j^t=\frac{\sum_{p=1}^{n}\gamma\left(m_{jp}+p_{jp}^{\alpha}\right)\left(1+\gamma\left\|z_p-v_j^{t-1}\right\|^2\right)^{-2}z_p}{\sum_{p=1}^{n}\gamma\left(m_{jp}+p_{jp}^{\alpha}\right)\left(1+\gamma\left\|z_p-v_j^{t-1}\right\|^2\right)^{-2}} \quad (4)$$

where '*t*' denotes the *t*[th] iteration.

3.3 Typicality

Using the required condition of the Lagrangian technique, the objective function is minimised, yielding the following generalised membership of typicality:

$$\Rightarrow p_{jp}=\frac{\left(\left[\left(1-C_w(z_j,v_p)\right)\right]\right)^{\frac{1}{-(\alpha-1)}}}{\sum_{l=1}^{n}\left(\left(1-C_w(z_l,v_j)\right)\right)^{\frac{1}{-(\alpha-1)}}} \quad (5)$$

3.4 Evaluation of C_j

To derive the updating equation for computing C_j the above proposed model is minimized with respect to C_j. Differentiating partially with respect to C_j, we get

$$C_j=\frac{\sum_{p=1}^{n}m_{jp}}{n} \quad \text{for all } j=1,2,\ldots,c \quad (6)$$

The steps of MPKFCW Procedure:

- *Set cluster number*
- *choose the centers of each cluster*
- *Compute the degree of membership through (3)*
- *Update the cluster center using (4)*
- *Estimate the typicality value using (5)*
- *Evaluation of C_j by (6)*
- *Repeat Step 3, 4, 5 & 6 till the process reaches the result*

4 Experimental Work

The efficacy and performance of the suggested fuzzy clustering approach were evaluated through a set of trials in this study. The proposed approach has been examined in experimental studies using the generated and TAE datasets [15]. This section first demonstrated the efficacy of the proposed technique using artificial data. The results of the Existed Method-1[5] and Existed Method-2 [16] are given in Figures 2(i) and 2(ii). The proposed fuzzy c-means approach influenced the effective allocation of data elements to accurate clusters with target functions derived by the Cauchy kernel based on distance measurements. The result of the proposed algorithm is given in Fig. 2(iii). This figure also shows that *MPKFCW* completely separates the two clusters.

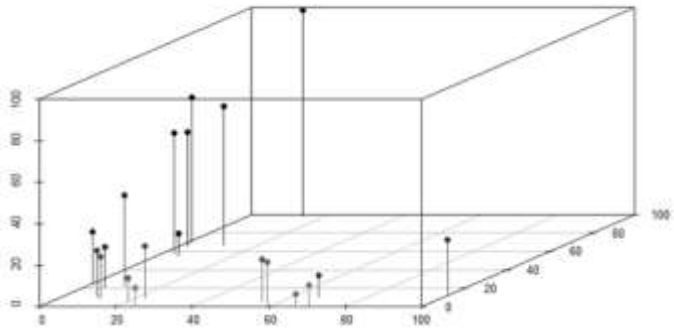
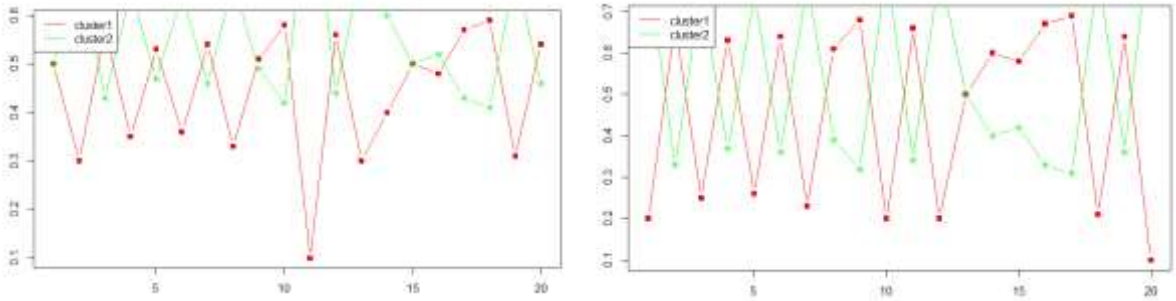
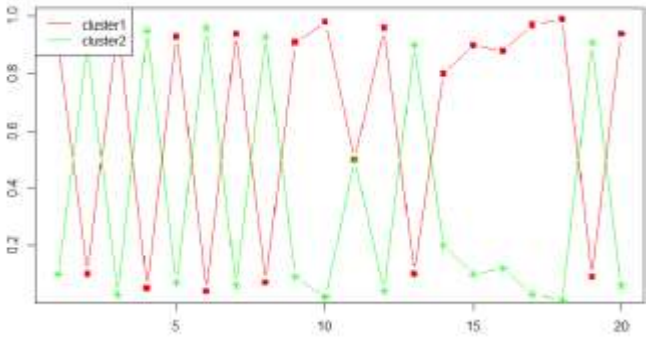


Fig:1 Artificial Data



(i) ***(ii)***



(iii)

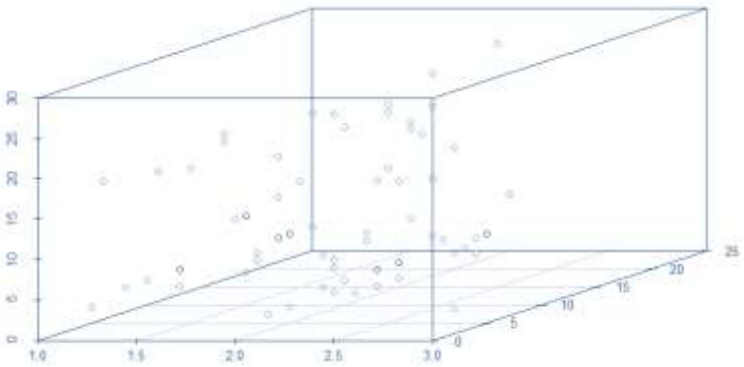


Figure 2. Memberships by (i) Existed Method-1 (ii) Existed Method-2 and (iii) Proposed Method

Fig 3: TAE Dataset

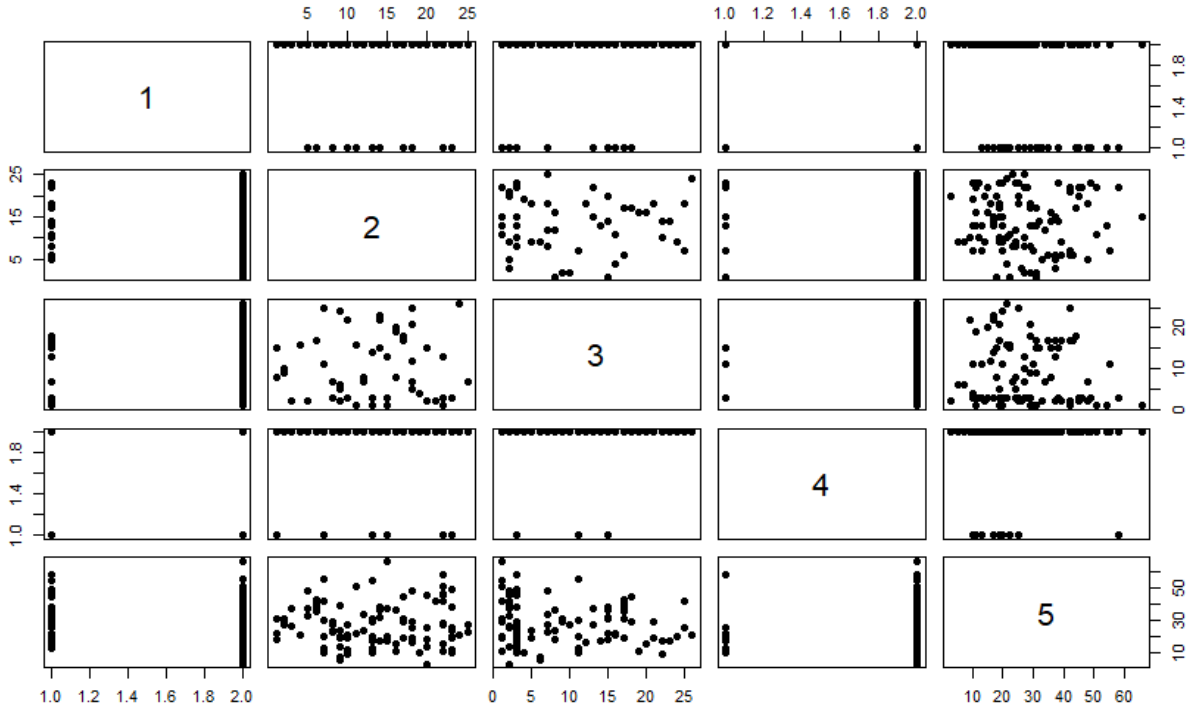
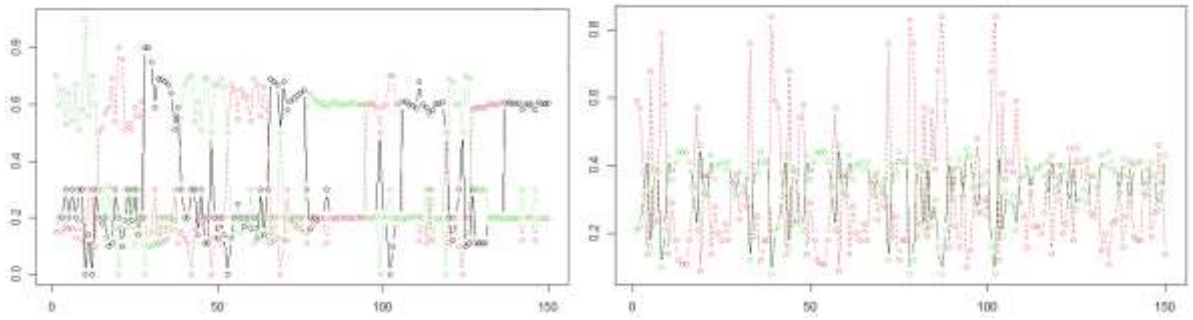
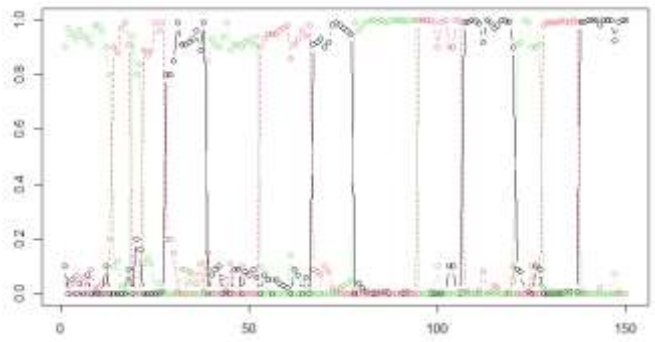


Fig 4: Correlation Plot of TAE Dataset



(i) ***(ii)***



(iii)

Fig 5. Memberships by (i) Existed Method-1 (ii) Existed Method-2 and (iii) Proposed Method

Table 1. Comparison of Algorithms

	Existed Method 1	*Existed Method 2*	*Proposed Method*
No. of Iterations	*11*	*8*	*5*
Clustering Accuracy	*81 %*	*89 %*	*97.2%*

This experiment splitting a TAE dataset into three clusters. The Teaching Assistant Evaluation dataset contains is depicted in Fig.3. This segment focuses specifically on the impact of the proposed method on TAE dataset. This part compares the outcomes of the proposed approach with the outcomes attained by existing approaches to demonstrating the usefulness of the proposed method in clustering TAE dataset. Fig 4. shows the correlation plot from a TAE dataset. The results of the existing clustering approaches are shown in Figures 5(i) and 5(ii). Due to the Cauchy kernel with weighted term, the proposed approach has been affected in efficiently allocating data items into accurate clusters. The outcome of the proposed method is shown in Fig. 5(iii), and this figure also shows that the proposed method completely separated the three clusters.

The Existed Methods entails more iterations to achieve clustering of the three subtypes of the TAE dataset. In addition, existing method-1 diminish the accuracy of the TAE database. Compared to the proposed approach used in this experiment, the accuracy values showed lower accuracy values for the existed algorithms. The proposed method groups the TAE dataset into three clusters, as shown in Table 1 and Figure 6, with good accuracy [2], a shorter running time, and fewer repetitions.

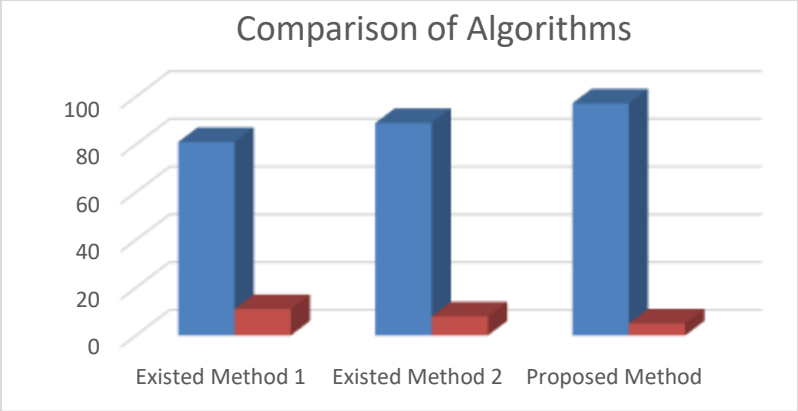


Fig 6: Comparison of Algorithms

5. Conclusion

This study introduces a Modified Cauchy kernel FPCM with weighted term based on membership functions, typical techniques, and kernel functions for cluster recognition in real world databases. This

work demonstrated experimental work on both artificial and TAE datasets to determine the effectiveness of the proposed approach. This study showed the superiority of the proposed algorithm in grouping similar expressions in a benchmark dataset by showing the number of iterations representing the accuracy of clustering, reiterations count, and suitably partitioned clusters.

References

1. *Abdellahoum Hamza et al., CSFCM: An improved fuzzy C-Means image segmentation algorithm using a cooperative approach, Expert Systems with Applications 166(3):114063*
2. *Abdulnassar et al., A Comprehensive Study on the Importance of the Elbow and the Silhouette Metrics in Cluster Count Prediction for Partition Cluster Models, Revista GEINTEC, Vol. 11 No. 4 (2021)*
3. *Adamyan, L., Efimov, K., Chen, C.Y. et al. Adaptive weights clustering of research papers. Digit Finance 2, 169–187 (2020).*
4. *Ankita Singh et al., Comparison Of K- Means and Fuzzy C- Means Algorithms, International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 5, May - 2013 ISSN: 2278-0181*
5. *Krishnamoorthy et al., A Comparative Study of Clustering Algorithm for Lung Cancer Data, International Journal of Scientific & Engineering Research, Volume 7, Issue 9, September-2016 1022 ISSN 2229-5518*
6. *Krishnapuram, R., and J.M. Keller. "A Possibilistic Approach to Clustering." IEEE Transactions on Fuzzy Systems 1, no. 2 (May 1993): 98–110.*
7. *Mapari et al., Study of Fuzzy Set Theory and Its Applications, IOSR Journal of Mathematics, Volume 12, Issue 4 Ver. II (Jul. - Aug.2016), PP 148-154*
8. *Pal et al., "A possibilistic fuzzy c-means clustering algorithm," IEEE TFS, vol. 13, no. 1, pp. 517–530, 2005*
9. *Pethalakshmi et al., Optimized K-Means and Fuzzy C-Means for MRI brain Image Segmentation, International Journal of Computational Intelligence and Informatics, Vol. 2: No. ISSN: 2349 – 6363*
10. *Preeti Jha et al., Apache Spark based kernelized fuzzy clustering framework for single nucleotide polymorphism sequence analysis, Computational Biology and Chemistry, Volume 92, June 2021, 107454*
11. *Rani Nooraeni et al., Fuzzy Centroid and Genetic Algorithms: Solutions for Numeric and Categorical Mixed Data Clustering, Procedia Computer Science, Volume 179, 2021, Pages 677-684*
12. *Shabtari et al., Analyzing PIMA Indian Diabetes Dataset through Data Mining Tool 'RapidMiner',* International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, 2021, pp. 560-574*
13. *Sneha et al., Advanced results in fuzzy sets and application in advanced materials, Materials Today: Proceedings, 2021*
14. *Seyyit et al, Increasing energy efficiency of rule-based fuzzy clustering algorithms using CLONALG-M for wireless sensor networks, Applied Soft Computing, Volume 109, September 2021, 107510*
15. *Tjen-Sien Lim e al., A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-three Old and New Classification Algorithms. Machine Learning 40, 203-228, 2000*
16. *Vanisri et al., An Efficient Fuzzy Possibilistic C-Means with Penalized and Compensated Constraints, Global Journal of Computer Science and Technology, 11(1), (2011)*

17. Xiaobin Zhi et al., Robust Local Feature Weighting Hard C-Means Clustering Algorithm, International Conference on Intelligent Science and Intelligent Data Engineering, 2011: Intelligent Science and Intelligent Data Engineering pp 591-598
18. Yanli Qi., Application of fuzzy clustering of massive scattered point cloud data in English vocabulary analysis, Microprocessors and Microsystems, Volume 81, March 2021, 103718
19. Zohaib Jan et al., Multiple strong and balanced cluster-based ensemble of deep learners, Pattern Recognition, Volume 107, 2020, 107420