ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol 19, Issue 1, 2025

An Advanced Fuzzy C-Means Approach for Effective Big Data Clustering

Dr.R Venkat[1], *Mr .K Venkat Tiru Gopal Reddy[2]

[1]Associate Professor, St. Peters's Engineering College, Secunderabad, Telangana – 500100

[2]Assistant Professor, St. Martin's Engineering College, Secunderabad, Telangana – 500100

*Corresponding Author

Email:kvenkatreddyit@smec.ac.in

Clustering emerged as powerful mechanism to analyze the massive data generated by modern applications; the main aim of it is to categorize the data into clusters where objects are grouped into the particular category. However there are various challenges while clustering the big data recently. Deep Learning has been powerful paradigm for big data analysis, this requires huge number of samples for training the model, which is time consuming and expensive. This can be avoided though fuzzy approach. In this research work, we design and develop an Improvised Fuzzy C-Means (IFCM)which comprises the encoder decoder Convolutional Neural Network (CNN) model and Fuzzy C-means(FCM) technique to enhance the clustering mechanism. Encoder decoder based CNN is used for learning feature and faster computation. In general FCM, we introduce a function which measure the distance between the cluster center and instance which helps in achieving the better clustering and later we introduce Optimized Encoder Decoder (OED) CNN model for improvising the performance and for faster computation. Further in order to evaluate the proposed mechanism, three distinctive data types namely Modified National Institute of Standards and Technology (MNIST), fashion MNIST and United States Postal Service (USPS) are used, also evaluation is carried out by considering the performance metric like Accuracy, Adjusted Rand Index (ARI) and Normalized Mutual Information(NMI).Moreover,comparativeanalysisisiscarriedoutoneachdatasetand
Comparative analysis shows that IFCM out performs the existing model.

Keywords:
Fuzzy C-Means (FCM), Convolutional Neural Network (CNN), improvised Fuzzy C-Means (IFCM)

1. INTRODUCTION

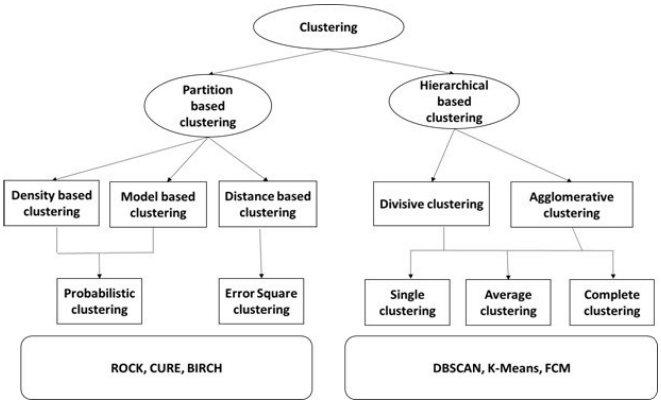
In recent times, enormous amount of data is being generated every day from various sources such as social media, satellites, sensors, mobile devices, computer simulations and business transaction. This data produces valuable information useful for business intelligence, forecasting, decision support, intensive data research. Walmart has nearly 2.5 peta bytes and Face book stores nearly 30 peta bytes of data, such huge data is known as Big Data; mining such big data is necessary to extract the desired information [1-3]. In general data are classified into the three types i.e. Structured, Semi-structured and Unstructured. Major part of the data portion is unstructured data which cannot be handled through traditional method. Big data can be defined through three distinctive parameters volume, velocity and variety [4]. Velocity describes the speed at which the data is exchanged, captured, and generated. Variety of data refers to type of data i.e. data is not always available in the structured form. It explains the complexities.

Clustering is unsupervised; essential for analyzing the data, partitions data into various subsets in particular way that similar data is clustered [5-6]. Clustering structure can be defined through the below equation, let's consider C as the cluster set and C_1, C_2etc be the clusters. Clustering is considered to be one of the machine learning mechanism.

$$C_1 \cap C_2 \cap C_2 \ldots\ldots \cap C_n = \emptyset \quad (1)$$

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 1, 2025

Big Data Clustering can be described through two aspects single and multiple machine clustering. Single aims for consolidating the data objects in accordance with the specific parameter [7]; based on the partition which divides the dataset into the single partition **Figure1.**Types of clustering mechanism



through the distance for points classification based on their similarities. However, the drawback is, it requires the pre-defined parameter which is non-deterministic [8-10]. Figure 1 shows different types of Clustering. Euclidean distance computes the minimum distance observed among the available cluster and assigned points [11]. Existing clustering algorithm has advantage of simple implementation whereas drawback of this approach is that it fails miserably to deal with large amount of data.

2. LITERATURE SURVEY

In this section we review several existing methodology; at first VAT[12] discuss clustering through dissimilarity matrix to achieve the modified matrix such that various cluster are displayed as the dark block through diagonally which is used in the dark matter halos, however this works only for the large cluster data. Moshtaghi et al. [13] developed an approach clustering by anomaly detection; here dendograms were used for the visual representation and applied for several taxonomy applications [14]. Similarly, Wilbiketal.
[15] proposed single linkage-based clustering for segmenting the time series based data to monitor the patient. The VAT commercial application was used for security [16], further it is observed that K-means promises to cluster the data efficiently. The advantage of using K-means is its applicability and simplicity in several fields; as a batch based algorithm, it comes with various limitation as it has poor initialization. In recent years, deep learning has been one of the major research areas; a supervised learning task that has gained satisfactory results in big data clustering [17-20];fails to deliver the result among the raw data and it affects the accuracy. Hence several rough based or fuzzy based approach is developed for handling the uncertainty in clustering. Dengetal. [17] developed a hierarchical approach which integrates the neural network and fuzzy logic for the robust clustering; here they minimize the vagueness. In literature [20], a fuzzy based CNN model was developed for the classification and clustering, in here at first CNN was applied to automate the feature extraction from given any input image and later FCM approach was used for clustering the data in defined feature space.

Rajeshet al. [21] developed an approach based on neural network with rough set based to cluster the data. Set theory approach was used for extracting the feature and then produced as input for the Feed Forward (FF)-neural network to cluster data. This is succeeded in handling the data quiet well; however these are mainly supervised learning approach and requires huge data for training and this further causes the time consumption. Further semi-supervised clustering was introduced to handle the clustering and classification [19][22-23]; Wu and Prasad [19] developed the restricted labeled data using the pseudo label. At first predicted label is used for clustering algorithm and pre-train neural network along with predicted labels. Predicted label helps in extracting the discriminating features; further ne-tune were introduced for adjusting the features from given pre-trained network for more beneficial to the clustering and classification. Tarvainen and Valpola [24] proposed semi-supervised learning named MT-model; MT-model averages the model weight for formatting the teacher model. MT-model was designed for the online learning and large dataset.

An efficient deep neural net work was developed[25];self- ensemble was introduced to form the predicting the unknown label through network training the various epochs. Moreover, the above two mentioned performs great

ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 1, 2025

on the general dataset; but it fails on achieving the better accuracy on the noisy sample and uncertain dataset..

3. PROPOSED METHODOLOGY

In this section, we develop Proposed Mechanism based on the CNN for enhancing the clustering mechanism. This is partitioned into various segments; at first, we learn about the general FCM and further we introduce a function parameter to compute the distance between the cluster center and instance. Later OED-CNN is introduced for improvisation in performance metrics. At last, both sub-mechanisms are integrated and presented as IFCM. In this section we discuss proposed model for big data representation. Let's define $Z \in T^{K_1 \times K_2 \times \ldots \times K_1}$as N-order multi dimensional array with size of $K_1 \times K_2 \times \ldots \times K_P$;multi-dimensional array presents different big data types such as unstructured data, structured data and semi-structured data and the character strings which is stored in the rational database.

Initialization

In general clustering approaches, objects are assigned to the single cluster. Fuzzy concept allows objects to belong to more than single cluster. In this research work we modify the concept of FCM algorithm.

Improvised Fuzzy C-Means approach

3.3.1 Function parameter

In this section, the function parameter is introduced for computing the distance between the instance and CC for better clustering as FCM faces huge drawback due to the distance. In Improvised FCM each instance is considered as the multidimensional array for capturing the correlation over various modalities. Moreover before deploying the FCM Optimized Encoder Decoder is applied for training the model, moreover to train the model Optimized Encoder Decoder is designed in the next section. Table 2 below is the modified FCM Algorithm.

Table2. ModifiedFCMAlgorithm

Input: Dataset, M, n, e *Output: optimized cluster member and membership vec*
Step1:Initializationofmembership matrix V
Step2:for k=1 to $\mathbb{M}$*do*
Step3:fork=1 to e do
Step4:clustercenterupdation $\eta_k = \sum_{l=1}^{p} w^l f_{kl}^{TD} / \sum_{(kl)}^{o} w^o_{l=1} \quad kl$
Step5:for k=1toedo
Step6:forl=1topdo
Step7: $w_{kl} = ((1+(_{\eta})^{\frac{f_{D_{(kl)}}}{i}^{-1/(o-1)^{-1}}}))$
Step8:end of for loop(step6)
Step9:end of for loop (step5)
Step10:endof for loop(step2)

Computational model

Computational model utilizes the CNN as the basic module for pre-training the parameters which are time consuming and highly computational. Further we design the optimized version to reduce the time overhead and the computational without compromising the parameters. The optimized Neural Network takes input as $Z \in T^{K_1 \times K_2 \ldots \ldots \times K_P}$and reconstruction of same is represented as $Z \in T^{K_1 \times K_2 \ldots \ldots \times K_P}$.

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 1, 2025

$$hid\_layer_{l1\ldots.lP} = enc(\psi)(\sum^{K1\ldots.KP} d^{(1)}_{l1\ldots\ldots lP} + Y^{(1)}_{\alpha k1\ldots.kP} Z_{k1\ldots\ldots kP}) \quad (6)$$

$$out\_layer_{k1\ldots kP} = dec(\psi)(\sum^{L1\ldots.LP} d^{(1)}_{k1\ldots\ldots kP} + Y^{(1)}_{\beta l1\ldots.lO} hid\_layer_{l1\ldots\ldots lO}) \quad (7)$$

In above equation,K_1 indicates the number of dimension whereasL_1indicatesthehiddenlayer,enc is encoder and decoder is decoder; further here we use sigmoid function in the encoding layer and decoding layer. Reconstruction objective

is given through the below equation. Eq. (8) is objective of the current research, this is reconstruction objective.

$$L_{Vencdec(\psi)} = [\frac{1}{o} \sum_{m=1}^{K1\times\ldots\ldots\times KOL1} \left(\sum_{s=1}^{LP} \sum_{l1=1}^{} \ldots \sum_{lO=1}^{} (Y^{(2)}_{sl1\ldots\ldots lO})^{2} \right)_{V} + (0.5(out\_layer_m - Z_m))I((out\_layer_m - Z_m)) + 0.5\zeta(\sum_{r=1}^{L1\times\ldots\ldots\times LnK1} \sum_{k1=1}^{KO} \ldots \sum_{jP=1}^{} Y^{(1)}_{rk1\ldots\ldots kO})^{2}] \quad (8)$$

Further back propagation is used for training the parameter.

Optimized Encoder Decoder CNN(OED-CNN)

OED-CNN is designed to minimize the time and computational overhead without affecting the performance. Optimized ANN comprises two hidden layer. OED-CNN is same as the encoder decoder based CNN except here we introduced ual approach for better training of model. Figure2 below is the OED-CNN Model.

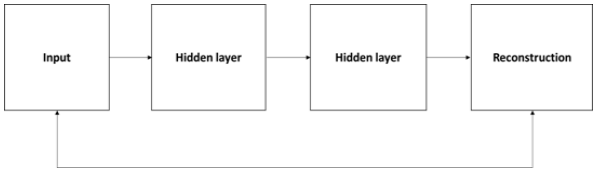


Figure2.OED-CNNModel

Further Table 3 provides the whole process of improvised FCM with OED-CNN model.

Table3. Improvised FCM with OED-CNN model

Input:$\mathbb{M}$, *dataset*
Step1:for edc=1tom do
Step2:forl_n*=1to*L_n*do* *Compute forward propagation using C means* *end for loop*

ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 1, 2025

Step3:form$_t$=1 to $M_t$do *Compute forward propagation for second layer using the forward propagation of first* *End for loop*
Step4:for k_p=1 to $K_p$doComputeoutputusingtheequation14 *End for loop*
Step5: if($L_{encdec}(\varphi)$>threshold)
Step6:form=1 to $K_p$do *Use the trainings amp le to formulate $^{(4)}_{k}$* *end for loop*
Step7:form$_t$=1toM_t(s=1,...,T)doUsing global training sample *End for loop*
Step8:forl_1=1to O do *Use parameter to compute $\tau^{(2)}k_1,k_2...k_O$* *end for loop*
Step9:fork_p=1toK_p(p=1,...,P)doComput e Δd *Form$_t$=1toM_t(t=1,...,T)doComp ute $\Delta Y^{(n)}$* *End for loop*
Step10:form$_t$=1toM_t(t=1,...,T)doCompu te Δd *For l_1=1,...,$L_{(n)}$* *doCompute $\Delta Y^{(n)}$* *End for loop*
Step11:forl_n=1,...,$L_n$doCompute Δd *For j_o=1toJ_p(p=1,...,P)doCompute ΔY* *End for loop* *End for loop*
Step12:updateparameters *$Y=Y-\alpha\Delta Y$˙* *$d=d-(\Delta d/n)\times\alpha$* *End if statement* *End for loop*

IFCM provides the better and faster clustering accuracy.

PERFORMANCE EVALUATION

Data set details

In this section, we provide a detailed description regarding the dataset; moreover three distinctive world dataset as MNIST, Fashion-MNIST and USPS; these dataset is considered for clustering. Fashion-MNIST is one of the popular fashion clothing dataset.

Comparison Algorithm

Fuzzy C-Means: This uses the membership matrix and update rule for clustering.

K-means: Here data can belong to one particular cluster. SEC: This is mainly based on the manifold learning.

MBKM: This algorithm is improvisation of K-means algorithm where mini-batch is used for minimizing the computational complexity.

DEC: This algorithm is mainly based on the deep learning, further this clustering model is based on the particular

designed distribution and abandons the decoder part.

IDEC: This is one of the deep clustering models; further this clustering model is based on the particular designed distribution and uses the reconstruction mechanism for regularizing the auto encoder.

Performance metrics

Normalized Mutual Information(NMI)

In general, mutual in form action is defined as them ensure of mutual dependence between two variables. NMI aka normalization of mutual information lies between 0 to 1, 0 indicates no mutual information and 1 indicates the perfect correlation. Higher NMI value indicates the better clustering model.

$$NMI=(H(E)+H(A))(H(E,A))^{-1} \quad (25)$$

Adjusted R and Index(ARI)

Rand Index is nothing but measure of similarity between two distinctive data clustering, Rand Index has value of range between 0 and 1, 0 indicates that two distinctive data clustering at any point and1 indicates that data clustering are absolute. Higher value of ARI indicates the higher efficiency of model.

$$\begin{aligned} &AverageRand\ index \\ &\quad =(Rand\_Index \\ &\quad -truenegative) \\ &\quad /(\max(Rand\_Index) \\ &\quad -E(Rand\_Index))^{-1} \end{aligned} \quad (26)$$

Accuracy

$$Clustering\_accuracy = P\left(\sum_{k=1}^{P} 1(A_k=\max(d_k))\right)^{-1} \quad (27)$$

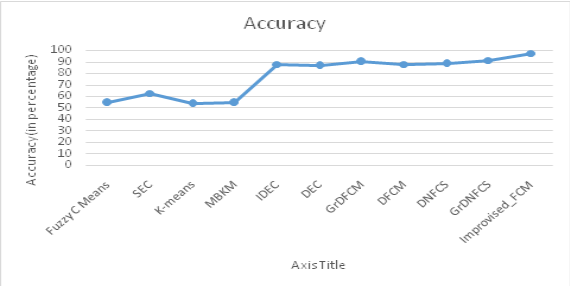
In the above equation, $d_k$indicates the clustering assignment.

Modified National Institute of Standards and Technology (MNIST) dataset

In this section, a comparative analysis of various method based on the three discussed metric is carried out. In here, it is observed that FCM achieves the very less accuracy of 54.68%, whereas other method like K-means and MBKM failsmiserablywithaccuracyof53.48%and54.43%.Further the other improvised methodology promises for better accuracy with 97.71% existing model achieving 91.45%. Similarly, in terms of ARI, FCM and K-means remains on the lower side with ARI value of 36.96% and 36.67%; other method like IDEC, DEC shows the marginal improvement with 88.01% and 86.53% respectively. In comparison with this entire model our model achieves 95.02%. Table 4 below is the performance metric comparison on MNIST dataset and accuracy graph is shown in the Figure 3.

Table4. Performance metric comparison on MNIST dataset

Clustering	Accuracy	ARI	NMI



INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 1, 2025

Methodologies			
Fuzzy C-Means	54.68	36.96	48.16
SEC[28]	62.73	48.59	60.38
K-means	53.48	36.67	49.99
MBKM[29]	54.43	36.85	44.82
IDEC[30]	88.01	83.25	86.38
DEC[31]	86.53	80.29	83.69
Gr DFCM	90.24	84.97	88.67
DFCM	88.17	83.37	86.54
DNFCS	88.26	83.44	86.65
Gr DNFCS[32]	91.45	86.26	90.74
Improvised_FCM	97.71	93.874	95.024

Figure3. Comparison of various existing model on MNIST dataset

United States Postal Service(USPS)

Further evaluation of improvised FCM is carried out considering the comparison metric as accuracy, ARI and NMI on USPS dataset; Table5presents the comparison. In here existing method like fuzzy C-means achieves decent accuracy of 66.34% and K means achieves 66.79%. Other existing method like DFCM and DNFCS shows some promising result with accuracy of 75.36% and 75.8% respectively.

Table5. PerformancemetriccomparisononUSPSdataset

Clustering Methodologies	Accuracy	ARI	NMI
Fuzzy C-Means	66.34	53.93	62
SEC	65.19	49.36	64.88
K-means	66.79	54.5	62.56
MBKM	62.87	51.05	59.93
IDEC	75.13	67.91	75.95
DEC	72.78	66.22	73.52
Gr DFCM	76.03	68.83	77.25
DFCM	75.36	68.15	76.36
DNFCS	75.8	68.77	76.96
Gr DNFCS	76.52	69.03	77.61
Improvised_FCM	95.12	85.01	89.01

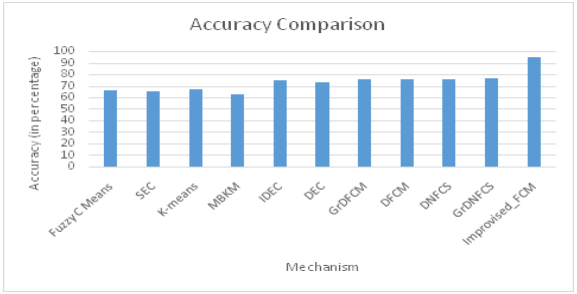


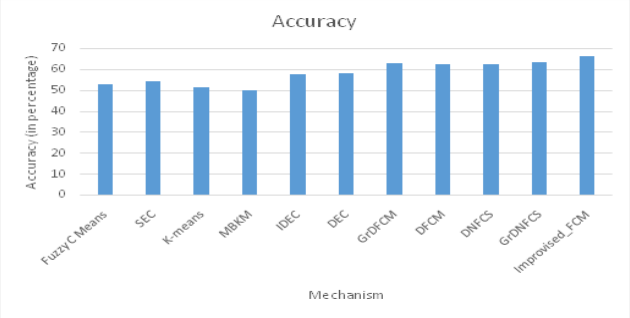
Figure4. Comparison of various existing model on USPS dataset

Fashion MNIST

In this sub-section comparative analysis is carried out on the Fashion MNIST dataset; it is one of them cost complicated dataset. Table 6 shows the comparison of various existing mechanism with proposed model in terms of accuracy, ARI and NMI. Moreover, Basic Fuzzy C-means achieves accuracy of 52.91% and K-means achieves accuracy of 51.07%. However other method like IDEC, DEC, DFCM achieves better accuracy but it stays on lower side; furthermore improvised FCM achieves decent accuracy of 66.2% in comparison with existing model of 63.51%. Similarly considering ARI as comparison metric, it is observed that Fuzzy C-means achieve ARI value of 36.44% and K-means achieves ARI value of 36.39%; other existing model gives decent improvisation with DFCM achieving 48.65% and existing model achieving 50.28%. Besides, in comparison with other existing

ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 1, 2025

model, Improvised FCM achievesdecentARIvalueof54.19%.Finally,NMI is considered as the comparison metric, where Fuzzy C-means achieves 51.59% and K means achieves 51.64%. Moreover, existing model achieves



66.09% whereas improvised FCM achieves 67.35%.

Figure5 below is comparison of various existing model on Fashion MNIST dataset.

Figure5. Comparison of various existing model on Fashion MNIST dataset

4. CONCLUSION

IFCM comprises the general FCM with additional function parameter for computing the distance between instance and CC; Further we introduce OED-CNN to enhance the performance metrics. Moreover optimized encoder decoder CNN helps in training the model in efficient and faster way; combined with fuzzy C-Means, IFCM possesses fine clustering model. Further to evaluate IFCM, three established machine learning datasets are considered i.e. MNIST, Fashion-MNIST and USPS. Also, detailed comparative analysis is carried out considering performance metric as accuracy, normalized mutual index and adjusted rand index; in each of these metric IFCM excels in comparison with various state-of-art techniques like FCM and K-means. In machine learning area, clustering is considered as novice mechanism for data analysis; although IFCM possesses great clustering mechanism with marginally growth in comparison with other exiting models. There are several other areas which need to be focused for real time data clustering.

REFERENCES

[1] Chen, C.L.P., Zhang, C.Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Inf. Sci. (Ny)., 275: 314-347. https://doi.org/10.1016/j.ins.2014.01.015

[2] Wang,X.K.,Yang,L.T.,Liu,H.Z.,Deen,M.J.(2017). A big data-as-a-service framework: State-of-the-art and perspectives. IEEE Trans. Big Data, 4(3): 325-340. https://doi.org/10.1109/TBDATA.2017.2757942

[3] Elkano, M.,Sanz, J.A.A.,Barrenechea, E.,Bustince,H., Galar, M. (2019). CFM-BD: A distributed ruleinduction algorithm for building Compact fuzzy models in big data classification problems. IEEE Trans. Fuzzy Syst., vol. 1. https://doi.org/10.1109/TFUZZ.2019.2900856

[4] Kumar,D.,Bezdek,J.C.,Palaniswami,M.,Rajasegarar, S., Leckie, C., Havens, T.C. (2016). A hybrid approach to clustering in big data. IEEE Transactions on Cybernetics, 46(10): 2372-2385. https://doi.org/10.1109/TCYB.2015.247741

[5] Deng,Y.,Ren,Z.Q.,Kong,Y.Y.,Bao,F.,Dai,Q.H. (2017). A hierarchical fused fuzzy deep neural network for data classification. IEEE Trans. Fuzzy Syst., 25(4): 1006-1012. https://doi.org/10.1109/TFUZZ.2016.2574915

[6] Blum, A.L., Langley, P. (1997). Selection of relevant features and examples in machine learning. Artif. Intell., 97(1-2):245-271.https://doi.org/10.1016/S0004-3702(97)00063-5

[7] Han, J., Kamber, M., Pei, J. (2011). Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers.
[8] Rokach, L. (2005). Clustering Methods. Data Mining and Knowledge Discovery Handbook. Springer, pp.331-352.
[9] Saxena, A., Pal, N.R., Vora, M. (2010). Evolutionary methods for unsupervised feature selection using Sammon‟s stress function. Fuzzy Information and Engineering, 2: 229-247. https://doi.org/10.1007/s12543-010-0047-4
[10] Jain, A.K. (2008). Data clustering: 50 years beyond k- means.PatternRecognitionLetters,31(8): 651-666.https://doi.org/10.1016/j.patrec.2009.09.011
[11] Estivill-Castro, V., Yang, J.H. (2000). Fast and robust general purpose clustering algorithms. In: Mizoguchi R., Slaney J. (eds) PRICAI 2000 Topics in Artificial Intelligence. PRICAI 2000. Lecture Notes in Computer Science, vol 1886. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-44533-1_24
[12] Bezdek, J.C., Hathaway, R.J. (2002). VAT: A tool for visual assessment of (cluster) tendency.In Proc. Int. JointConf. Neural Netw.(IJCNN),Honolulu, HI, USA, pp. 2225–2230.https://doi.org/10.1109/IJCNN.2002.1007487