IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 34, No. 6, December, 2020, pp. 701-708
Journal homepage: http://iieta.org/journals/ria

Big Data Clustering Using Improvised Fuzzy C-Means Clustering

Venkat Rayala*, Satyanarayan Reddy Kalli

Research Centre, Department of CSE, Cambridge Institute of Technology, Affiliated to Visvesvaraya Technological University (VTU)-Belgaum, Bengaluru 560036, India

Corresponding Author Email: rayalavenkat534@gmail.com

https://doi.org/10.18280/ria.340604

Received: 24 October 2020
Accepted: 30 November 2020

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

ABSTRACT

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, comparative analysis is carried out on each dataset and comparative analysis shows that IFCM outperforms the existing model.

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 petabytes and Facebook stores nearly 30 petabytes 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_2 etc be the clusters. Clustering is considered to be one of the machine learning mechanism.

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

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 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.

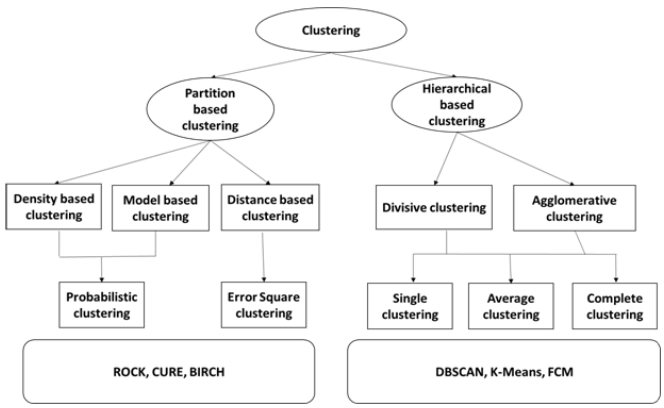

Figure 1. Types of clustering mechanism

1.1 Motivation and contribution of research work

Clustering mechanism requires the absolute weights for weighted distance in accordance with requirement; the existing clustering approach uses assigned weights randomly. Moreover, fuzzy and neural network aims at exploiting the knowledge processing; through the survey it is observed that very few researchers have considered deep learning concept and none of them have considered encoder decoder based CNN. Motivated by this phenomenon we have developed IFCM which comprises encoder decoder CNN with FCM for better performance metrics.

In this work an IFCM is developed for efficient and high accuracy intended big data clustering.

IFCM comprises of distinctive framework namely FCM, FCM with function parameter and OED-CNN.

- FCM approach, we introduce a function which measure the distance between the Cluster Center (CC) and instance which helps in achieving the better clustering. Improvised FCM has Dual CNN for handling the big data in efficient manner.
- Later we introduce OED- CNN model for enhancing the performance metrics and faster computation.

IFCM is evaluated through considering the well-established data such as MNIST, fashion-MNIST, USPS; further comparative analysis is carried out with the existing model. IFCM performs better than several existing model.

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, Wilbik et al. [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. Deng et al. [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.

Rajesh et 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 network 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 on the general dataset; but it fails on achieving the better accuracy on the noisy sample and uncertain dataset. Apart from this research work like the literatures [26, 27] focused on discussing advantage of FCM algorithm over the other clustering technique. Considering the above existing methodology, we observe that all this clustering mechanism faces problem of computational time, absolute clustering, and performance metrics. However, through the research gap analysis it was observed that FCM possesses a great potential in comparison with the other existing technique like k-means or other traditional method. Most of the existing model follows hard clustering which categorizes the object into one category whereas FCM is soft clustering. However, FCM takes more computational time and fails in metrics, hence in the next section we design and develop improvised FCM which is based on the CNN mechanism.

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 multidimensional 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.

3.1 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.

3.2 System model and general FCM algorithm

FCM algorithm operates through assigning the membership for each data point to correspondent CC based on the data point and cluster distance; the main advantage of FCM is that it provides the outstanding result in case of overlapped data and also it assigns the data point to more than one cluster. However apart from computational time and accuracy it requires a greater number of iteration and Euclidean distance is used which measures the weight in unequal manner. Hence this can be reduced through encoder decoder based CNN.

Let us consider the dataset $Z = \{z_1, z_2, \ldots, z_q\}$ with cluster set $X = \{x_1, x_2, \ldots, x_p\}$ and membership set $W = \{w_{kl} | 1 \le k \le e, 1 \le l \le p\}$; further considering these three FCM can be formulated. In general, the idea of proposed mechanism is to integrate the IFCM with double neural network. Further we develop an optimized auto-encoder for training the for instance.

$$\begin{aligned} min: & \sum_{k=1}^{e}\sum_{l=1}^{p} w_{kl}^{o}\|z_l - x_k\|^2 \\ & \sum_{l=1}^{e} w_{kl} = 1, w_{kl} \ge 0 \end{aligned} \tag{2}$$

To avoid false clustering, we develop modified-FCM, Modified FCM in the below equation.

$$\begin{aligned} L_o(W,X) = & \sum_{k=1}^{e}\eta_i \sum_{k=1}^{p}(1-u_{kl}^{o})^{o} \\ & + \sum_{k=1}^{e}\sum_{k=1}^{p} w_{kl}^{m}\|z_l - z_i\|^2 \end{aligned} \tag{3}$$

Hence optimizing the equation helps in updating the membership matrix as well as cluster centers and given as:

$$x_k = \sum_{l=1}^{p} w_{kl}^{o}\, z_l / \sum_{l=1}^{p} w_{kl} \tag{4}$$

Membership matrix:

$$w_{kl} = \left(1 + \left(\frac{e_{kl}}{\eta_k}\right)^{-1/(o-1)}\right)^{-1} \tag{5}$$

In the above equation, e_{kl} indicates the distance between the cluster and membership matrix. Table 1 below is the General FCM Algorithm.

Table 1. General FCM Algorithm

Input: dataset, Max
Output: optimized cluster member and membership vector
Step1: Initialization of membership_matrix
Step2: for k =1 to e do
Step3: cluster center updation
Step4: updating the fuzzy constant
Step5: for =1 to e do
Step6: for l=1 to e do
Step7: cluster center updation
Step8: end for loop(step7)
Step9: end of for loop (step6)
Step10: end of for loop(step2)

3.3 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.

Table 2. Modified FCM Algorithm

Input: Dataset, M, n, e
Output: optimized cluster member and membership vec
Step1: Initialization of membership matrix V
Step2: for k=1 to $\mathbb{M}$ do
Step3: for k=1 to e do
Step4: cluster center updation $\eta_k = \sum_{l=1}^{p} w_{kl}^{l} f_{TD_{(kl)}} / \sum_{l=1}^{o} w_{kl}^{o}$
Step5: for k= 1 to e do
Step6: for l= 1 to p do
Step7: $w_{kl} = \left(\left(1 + \left(\frac{f_{TD_{(kl)}}}{\eta_i}\right)^{-1/(o-1)}\right)^{-1}\right)$
Step8: end of for loop(step6)
Step9: end of for loop (step5)
Step10: end of for loop (step2)

3.3.2 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 $Z \in T^{K_1 \times K_2 \ldots \times K_P}$ and reconstruction of same is represented as $Z \in T^{K_1 \times K_2 \ldots \times K_P}$.

$$\begin{aligned} hid\_layer_{l_1 \ldots l_P} = enc(\psi)(& \sum_{k_1 \ldots k_P}^{K_1 \ldots K_P} d_{l_1 \ldots l_P}^{(1)} \\ & + Y_{\alpha k_1 \ldots k_P}^{(1)} Z_{k_1 \ldots k_P}) \end{aligned} \tag{6}$$

$$\begin{aligned} out\_layer_{k_1 \ldots k_P} = dec(\psi)(& \sum_{l_1 \ldots l_O}^{L_1 \ldots L_P} d_{k_1 \ldots k_P}^{(1)} \\ & + Y_{\beta l_1 \ldots l_O}^{(1)} hid\_layer_{l_1 \ldots l_O}) \end{aligned} \tag{7}$$

In above equation, K_1 indicates the number of dimension whereas L_1 indicates the hidden layer, enc is encoder and dec 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.

$$L_{V\,encdec(\psi)} = \left[\frac{1}{o}\sum_{m=1}^{o}\left(\sum_{s=1}^{K_1\times\ldots\ldots\times K_O}\sum_{l_1=1}^{L_1}\ldots\sum_{l_O=1}^{L_P}\left(Y^{(2)}_{sl_1\ldots\ldots l_O}\right)\right)^2 + \left(0.5(\text{out\_layer}_m - Z_m)\right)^V I\left((\text{out\_layer}_m - Z_m)\right) + 0.5\zeta\left(\sum_{r=1}^{L_1\times\ldots\ldots\times L_n}\sum_{k_1=1}^{K_1}\ldots\sum_{j_P=1}^{K_O}Y^{(1)}_{rk_1\ldots\ldots k_O}\right)^2\right] \quad (8)$$

Further back propagation is used for training the parameter.

3.3.3 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 introduce dual approach for better training of model. Figure 2 below is the OED-CNN Model.

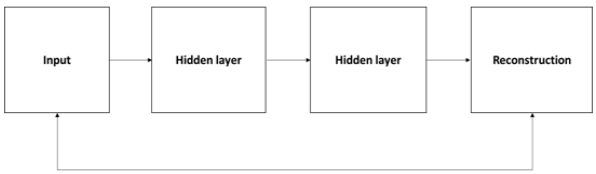

Figure 2. OED-CNN Model

OED-CNN takes an input as $Z \in T^{K_1\times K_2\ldots\ldots\times K_P}$; further operation at both the hidden layer is given through the respective equation.

$$hidden\_layer_{1l1\ldots\ldots lO}^{K_1\ldots\ldots K_P} = enc(\psi)\left(\sum_{k_1\ldots\ldots k_P} Y^{2(1)}_{\alpha j_1\ldots\ldots j_P} Z_{j_1\ldots\ldots j_P} + d^{(1)}_{l1\ldots\ldots lO}\right) \quad (9)$$

In above equation, hiddden_layer_1 is the first layerand $\text{hiddeen\_layer}_2 \in T^{M_1\times M_2\ldots\ldots\times M_T}$.

$$hidden\_layer_{2m_1\ldots m_T} = enc(\psi)\left(\sum_{l_1\ldots l_P}^{L_1\ldots L_P} Y^{(1)}_{\alpha l_1\ldots l_P} hidden\_layer_{l_1\ldots l_P} + d^{(1)}_{l1\ldots lO}\right) \quad (10)$$

Similarly, $\text{hidden\_layer}_1 \in T^{L_1\times L_2\times\ldots\ldots\times L_N}$ indicates the second layer and $\text{Hidden\_layer}_2 \in T^{M_1\times M_2\ldots\ldots\times M_O}$; in both equations $enc(\psi)$ indicates the encoder. And the output is given through Y and represented in the below equation; here $dec(\psi)$ indicates the decoder.

$$output\_layer_{k1\ldots kO} = dec(\psi)\left(\sum_{l_1\ldots l_O}^{M_1\ldots M_T} d^{(2)}_{k1\ldots kO} + Y^{(1)}_{\beta k_1\ldots k_O} hidden\_layer_{2m_1\ldots m_T}\right) \quad (11)$$

Moreover, in the optimized ANN training model, Rectifier unit is used in the encoder function as the activation function and it is given as:

$$h'(z) = \begin{cases} 1 & z > 0 \\ 0 & z \le 0 \end{cases} \quad (12)$$

Further parameters are trained and reconstruction function is given as:

$$L_{ATAE}(\psi) = 0.5(a-z)^T I(a-z) \quad (13)$$

In the above equation, z and a are the vector; Idenotes the coefficient. Further the reconstruction function on the given m training sample is given as:

$$L_{GRF}\ (\psi) = \left[0.5\sum_{k=1}^{m} 0.5\zeta\left(Y^{(1)2} + Y^{(2)2} + Y^{(3)2}\right)^2 + 0.5\,(a-z)^T I(a-z)\right] \quad (14)$$

Later, back propagation is applied for computing the δY .

$$Y = Y\ \ - \varphi\left(\frac{1}{o}\sum_{k=1}^{o}\varphi Y + \delta Y_k\right) \quad (15)$$

Similarly back propagation is applied for computing Δd. Eq. (16) is the back propagation is applied for computing Δd.

$$\Delta d = d\ \ - \varphi\left(\frac{1}{o}\sum_{k=1}^{o}\delta d_k\right) \quad (16)$$

In the above equation φ indicates the rate of learning. Moreover in order to compute the derivative in case of each sample, forward propagation is applied for input and output value computation. Eq. (17) is Forward Propagation for input, Eq. (18) is the Forward Propagation of output.

$$\tau_k^{(4)} = \left(\sum_{l=1}^{K_1\times\ldots\ldots\times K_o} i_{kl}\left(z_k^{(3)} - a_k\right)\right).h'\left(b_k^{(4)}\right) \quad (17)$$

$$\tau^{(3)}_{m_1m_2\ldots m_T} = \left(\sum_{l=1}^{K_1\times\ldots\ldots\times K_o} Y^{(3)}_{kl_{1\ldots m_T}}\ .\tau_k^{(4)}\right).h'\left(b_k^{(4)}\right) \quad (18)$$

In above equation, $\tau^{(4)}_{l_1,l_2\ldots.l_T}$ and $\tau^{(3)}_{l1,l2\ldots,lO}$ are input and output values; error is computed for each neuron through below equation.

$$\tau^{(2)}_{m_1m_2\ldots m_T} = \frac{\partial L^{(4)}_{Genecdec}(\varphi)}{\partial b^{(3)}_{l_1\ldots l_P}} \quad (19)$$

Further partial derivative of $\frac{\partial b^{(4)}}{\partial Y^{(n)}}$ is computed by considering $n = 1,2$ and 3.

$$c^{(3)}_{m_1m_2\ldots m_T} = \frac{\partial b^{(4)}_{k_1\ldots k_O}}{\partial Y^{(3)}_{l_1\ldots l_S}} \quad (20)$$

$$c^{(2)}_{m_1 m_2 \ldots m_T} = \frac{\partial b^{(3)}_{k_1 \ldots k_O}}{\partial Y^{(2)}_{l_1 \ldots l_S}} \quad (21)$$

$$c^{(1)}_{m_1 m_2 \ldots m_T} = \frac{\partial b^{(2)}_{j_1 \ldots j_O}}{\partial Y^{(1)}_{k_1 \ldots k_R}} \quad (22)$$

Considering the chain rule we compute the derivatives of δY and Δd.

$$\Delta Y^{(m)} = \frac{\partial L_{Gencdec^{(\psi)}}}{\partial b^{(n+1)}} . \frac{\partial L^{(n+1)}}{\partial Y^{(n+1)}} \quad (23)$$

$$\Delta d^{(m)} = \varsigma^{(m+1)} \quad (24)$$

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

Table 3. Improvised FCM with OED-CNN model

Input: $\mathbb{M}$, dataset
Step1: for edc= 1 to m do
Step2: for l_n=1 to L_n do *Compute forward propagation using C means* *end for loop*
Step3: for m_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 do *Compute output using the equation 14* *End for loop*
Step5: if ($L_{encdec}(\varphi)$>threshold)
Step6: for m=1 to K_p do *Use the training sample to formulate $\tau_k^{(4)}$* *end or loop*
Step7: for m_t=1 to M_t (s=1,…,T) do *Using global training sample* *End for loop*
Step8:for l_1=1 to O do *Use parameter to compute $\tau^{(2)}k_1, k_2 \ldots k_O$* *end for loop*
Step9: for k_p=1 to K_p (p=1,…,P)do *Compute Δd* *Form$_t$=1 to M_t (t=1,…,T) do* *Compute $\Delta Y^{(n)}$* *End for loop*
Step10: for m_t=1 to M_t (t=1,…,T) do *Compute Δd* *For l_1=1,…,$L_{(n)}$ do* *Compute $\Delta Y^{(n)}$* *End for loop*
Step11: for l_n=1,…,L_n do *Compute Δd* *for j_o=1 to J_p (p=1,…,P) do* *Compute ΔY* *End for loop* *End for loop*
Step12: update parameters *Y=Y-α$\Delta Y^{\cdot}$* *d=d-(Δd/n)×α* *End if statement* *End for loop*

IFCM provides the better and faster clustering accuracy.

4. PERFORMANCE EVALUATION

In this section, the proposed mechanism is evaluated on real dataset for clustering; further comparative analysis is carried out. In order to evaluate the mechanism ideal system configuration of i7 processor packed with 2GB Nvidia graphics and 8GB RAM; further python is used as the programming language along with various machine learning libraries.

4.1 Dataset 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.

4.2 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.

4.3 Performance metrics

4.3.1 Normalized Mutual Information (NMI)

In general, mutual information is defined as the measure 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.

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

4.3.2 Adjusted Rand 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 and 1 indicates that data clustering is absolute. Higher value of ARI indicates the higher efficiency of model.

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

4.3.3 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.

4.3.4 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 fails miserably with accuracy of 53.48% and 54.43%. 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. Moreover, existing model achieves ARI value of 86.26%; however in comparison with that Improvised model achieves massive ARI of 93.87%. Furthermore, considering the NMI as metrics method like FCM and K-means achieves 48.16% and 49.99% respectively; improvising this existing model achieves NMI value of 90.74%. 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.

Table 4. Performance metric comparison on MNIST dataset

Clustering Methodologies	Accuracy	ARI	NMI
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
GrDFCM	90.24	84.97	88.67
DFCM	88.17	83.37	86.54
DNFCS	88.26	83.44	86.65
GrDNFCS[32]	91.45	86.26	90.74
Improvised_FCM	97.71	93.874	95.024

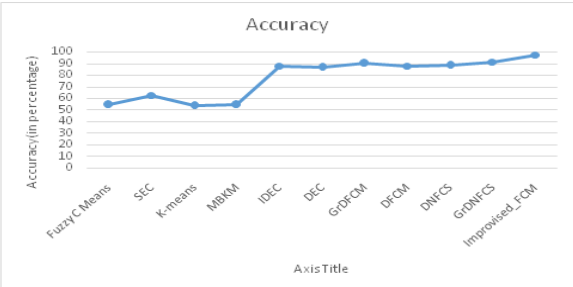

Figure 3. Comparison of various existing model on MNIST dataset

4.3.5 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; Table 5 presents 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. Moreover existing model i.e. GRDNFCS achieves 76.52% whereas IFCM achieves massive accuracy of 95.12%. Further considering the ARI metric FCM and K-means achieves ARI of 53.93% and 54.5%; other model like DFCM and DNFCS shows the decent improvisation with ARI of 68.15 and 68.77% respectively. In comparison with all these existing mechanism, Improvised FCM achieves 85.01%. At last, considering the NMI metric, Fuzzy C-Means achieves NMI of 68% and 64.88%; further DFCM and DNFCS shows promising with NMI of 76.36% and 76.96%. Moreover, in comparison with all these method and existing model, proposed model achieves highest NMI of 89.01%. Figure 4 shows the comparison graph in terms of accuracy.

Table 5. Performance metric comparison on USPS dataset

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
GrDFCM	76.03	68.83	77.25
DFCM	75.36	68.15	76.36
DNFCS	75.8	68.77	76.96
GrDNFCS	76.52	69.03	77.61
Improvised_FCM	95.12	85.01	89.01

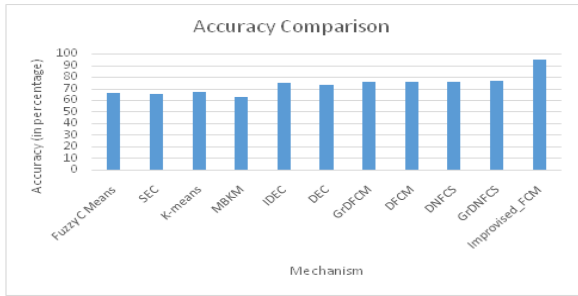

Figure 4. Comparison of various existing model on USPS dataset

4.3.6 Fashion MNIST

In this sub-section comparative analysis is carried out on the Fashion MNIST dataset; it is one of the most 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 model, Improvised FCM achieves decent ARI value of 54.19%. 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%.

Figure 5 below is comparison of various existing model on Fashion MNIST dataset.

Table 6. Performance Metric Comparison on Fasion MNIST dataset

Clustering Methodologies	Accuracy	ARI	NMI
Fuzzy C-Means	52.91	36.44	51.59
SEC	54.24	38.44	55.8
K-means	51.07	36.39	51.64
MBKM	50	34.5	50.03
IDEC	57.64	44.09	60.13
DEC	57.81	45.71	62.83
GrDFCM	62.78	50.14	65.78
DFCM	62.29	48.65	64.54
DNFCS	62.5	49.91	65.67
GrDNFCS	63.51	50.28	66.09
Improvised_FCM	66.2	54.19	67.35

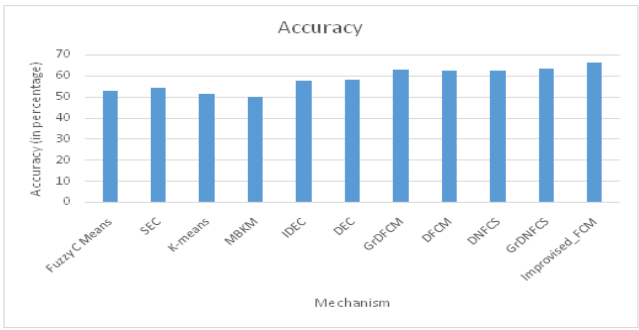

Figure 5. Comparison of various existing model on Fashion MNIST dataset

5. 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 rule induction 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. Pattern Recognition Letters, 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. Joint Conf. Neural Netw. (IJCNN), Honolulu, HI, USA, pp. 2225-2230. https://doi.org/10.1109/IJCNN.2002.1007487

[13] Moshtaghi, M., Havens, T.C., Bezdek, J.C., Park, L., Leckie, C., Rajasegarar, S., Keller, J.M., Palaniswami, M. (2011). Clustering ellipses for anomaly detection. Pattern Recognit., 44(1): 55-69. https://doi.org/10.1016/j.patcog.2010.07.024

[14] Sneath, P.H.A., Sokal, R.R. (1973). Numerical Taxonomy-The Principles and Practice of Numerical Classification. San Francisco, CA, USA: W.H. Freeman& Co.

[15] Wilbik, A., Keller, J.M., Bezdek, J.C. (2013). Linguistic prototypes for data from eldercare residents. IEEE Trans. Fuzzy Syst., 22(1): 110-123. https://doi.org/10.1109/TFUZZ.2013.2249517

[16] Zhang, D., Ramamohanarao, K., Versteeg, S., Zhang, R. (2009). RoleVAT: Visual assessment of practical need

for role based access control. in Proc. Conf. Comput. Security Appl., Honolulu, HI, USA, pp. 13-22. https://doi.org/10.1109/ACSAC.2009.11

[17] 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

[18] Riaz, S., Arshad, A., Jiao, L.C. (2018). Fuzzy rough C-mean based unsupervised CNN clustering for large-scale image data. Appl. Sci., 8(10): 1-20. https://doi.org/10.3390/app8101869

[19] Wu, H., Prasad, S. (2018). Semi-supervised deep learning using pseudo labels for hyperspectral image classification. IEEE Trans. Image Process., 27(3): 1259-1270. https://doi.org/10.1109/TIP.2017.2772836

[20] Yeganejou, M., Dick, S. (2018). Classification via deep fuzzy c-means clustering. 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Rio de Janeiro, pp. 1-6. https://doi.org/10.1109/FUZZ-IEEE.2018.8491461

[21] Rajesh, T., Malar, R.S.M. (2013). Rough set theory and feed forward neural network based brain tumor detection in magnetic resonance images. International Conference on Advanced Nanomaterials & Emerging Engineering Technologies, Chennai, pp. 240-244. https://doi.org/10.1109/ICANMEET.2013.6609287

[22] Kuznietsov, Y., Stuckler, J., Leibe, B. (2017). Semi-supervised deep learning for monocular depth map prediction.In Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 6647-6655.

[23] Zhou, S.S., Chen, Q.C., Wang, X.L. (2014). Fuzzy deep belief networks for semi-supervised sentiment classification. Neurocomputing, 131: 312-322. https://doi.org/10.1016/j.neucom.2013.10.011

[24] Tarvainen, A., Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In Proc. Adv. Neural Inf. Process. Syst., pp. 1195-1204.

[25] Laine, S., Aila, R. (2016). Temporal ensembling for semi-supervised learning. [Online]. Available: https://arxiv.org/abs/1610.02242

[26] Venkat, R., Reddy, K.S. (2019). Dealing big data using fuzzy c-means (FCM) clustering and optimizing with gravitational search algorithm (GSA). 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, pp. 465-467. https://doi.org/10.1109/ICOEI.2019.8862673

[27] Venkat, R., Reddy, K.S. (2020). Clustering of huge data with fuzzy c-means and applying gravitational search algorithm for optimization. International Journal of Recent Technology and Engineering (IJRTE), 8(5): 3206-3209. https://doi.org/10.35940/ijrte.D9130.018520

[28] Nie, F.P., Xu, D., Tsang, I.W., Zhang, C.S. (2009). Spectral embedded clustering. IJCAI, pp. 1181-1186.

[29] Sculley, D. (2010). Web-scale k-means clustering. Proceedings of the 19th international conference on World wide web. ACM, pp. 1177-1178. https://doi.org/10.1145/1772690.1772862

[30] Xie, J.Y., Girshick, R., Farhadi, A. (2016). Unsupervised deep embedding for clustering analysis. International conference on machine learning, 48: 478-487.

[31] Guo, X.F., Gao, L., Liu, X.W., Yin, J.P. (2017). Improved deep embedded clustering with local structure preservation. International Joint Conference on Artificial Intelligence (IJCAI-17), pp. 1753-1759. https://doi.org/10.24963/ijcai.2017/243

[32] Feng, Q.F., Chen, L., Chen, C.L.P., Guo, L. (2020). Deep fuzzy clustering-A representation learning approach. IEEE Transactions on Fuzzy Systems, 28(7): 1420-1433. https://doi.org/10.1109/TFUZZ.2020.2966173