Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 6, July, 2021: 8655- 8664

Research Article

# A Deep Conceptual Incremental learning Based High Dimensional Data Clustering model- A Deep Learning Approach

S. Praveen[1], Dr. R. Priya, MCA, M.Phil, Ph.D[2]

## Abstract

Clustering is one of the important and vital tasks in the natural language processing which is becoming familiar in the data based application domains. Performance of machine learning clustering models depends on the quality of data or learning representation. Machine learning algorithm plays an important role in high dimensional data clustering. However the algorithms are suffering with low accuracy and distribution of the data points. Deep learning approaches are currently explored for solving these challenges and for better data representation for clustering. In this paper, a novel deep learning approach named as Deep Conceptual Incremental Clustering has been proposed for analysing unstructured and high dimensional data. It implies the autoencoder model. It is effective in transform learning from high dimensional to low dimensional feature space to extract the concept specific features on the distribution of the data. It determines the salient features on ensuring the minimum reconstruction error. Architecture is composed of multiple layer to sparse representation and to eliminate the over fitting. Further all parameters are fine tuned with respect to certain criterion which considered as Loss function and encoder function. The encoder function is used to map the data points into latent representations. Finally it is helpful to find a better initialization of the parameters. Extensive experiments have been conducted on real datasets to compare proposed model with several state-of-the-art approaches. The experimental results show that Deep Conceptual Incremental clustering can achieve both effectiveness and good scalability on high dimensional data.

***Keywords:*** High Dimensional Data Clustering, Deep Learning, Data Distribution, Unstructured data, Reconstruction Error

## 1. Introduction

Clustering is a fundamental unsupervised learning representation employed in exploratory data mining applications. The primary objective of data clustering is to segment the distributed data points into more clusters based on similarity measures of the data points (e.g., Euclidean distance). However machine learning based clustering methods have been presented to distributed data clustering [1]. Traditional data clustering methods based on machine learning techniques usually have less clustering performance on high-dimensional data [2], especially due

[1] Research Scholar, Department of Computer Science Sree Narayana Guru College, Coimbatore
[2]Associate Professor and Head, Department of Computer Science Sree Narayana Guru College, Coimbatore

to the inefficiency of data similarity measures used by those methods. Further these data clustering methods generally suffer from curse of dimensionality, data sparsity and computational complexity on large-scale datasets [3][4]. To tackle those implications, dimensionality reduction and linear feature transformation methods such as Linear Discriminant Analysis(LDA) [5] such as Principal component analysis (PCA) [6] and non-linear feature transformation methods such as kernel methods [7] and spectral methods [8] have been extensively studied to map the raw data into a new feature space. Finally, a highly complex latent structure of distributed data also faces a lot of challenges in terms of the effectiveness on employing traditional clustering methods. Above mentioned challenges of traditional clustering techniques using machine learning has led to the development of deep learning mechanism to handle complex data representation in streaming applications, learning representation of the deep learning model can be used to transform the data into more clustering-friendly data distributions in sophisticated.

Deep learning model has distinct ability to discover a good representation of data. Especially jointly optimizing of the deep neural network with unsupervised clustering algorithm is becoming an active research field. Deep learning architecture for clustering has been employed using neural network for better representation of high dimensional data. In this article, Deep Conceptual Incremental Clustering has been proposed a deep learning approach for analysing unstructured and high dimensional data. Approach uses auto encoder category to obtain the feasible feature space containing concept specific features. The Auto encoder model provides non linear mapping function to construct the features with loss criteria to include the latent representation. Eventually, learning non-linear mappings allows transforming input data into more clustering-friendly representations in which the data is mapped into a lower-dimensional feature space

The Remaining paper is organized as follows, related work are described in section 2, the architecture of the proposed deep concept specific clustering approach is described in section 3 and experimental results and effectiveness of the proposed system is demonstrated in section 4 using dataset along performance comparison with state of arts approaches on various metric has been explained. Finally paper is concluded in section 5.

## 2. Related Work

In this section, data clustering model using machine learning approaches has been examined in details on basis of architectures for feature representations and similarity measures of the clustered data points. Each of those machine learning techniques which follows some kind of better performance effectiveness on the evaluation of the model has been represented in detail and few which performs nearly equivalent to the proposed model is described as follows

### 2.1.Ensemble based data Clustering

In this method, Ensemble based data clustering has been described for high dimensional data clustering. It partitions a given set of data points in a multidimensional space into clusters such that the points within a cluster are more similar to each other. Ensemble Clustering is represented in combination with dimension reduction techniques and feature selection and projection pursuit. It is considered as error-driven representativeness capture time-

changing concepts on the features. Finally the association learning has been incorporated for distributed data clustering [9].

### 2.2.Deep Adaptive Fuzzy Clustering

In this method, deep representation learning for distributed data clustering has been analysed. Learning architectures uses the deep adaptive fuzzy clustering to provide soft partition of data clusters on the highly complex structure in the streaming applications. Initially data reconstruction of streaming data composed of original data is transformed into meaningful space. The meaningful data is transformed into feature space using the word embedding process on the deep learning architecture. Word embedding process is a learnt representation of the text or sentence towards segmenting into feature vector containing words, characters and N-grams of words. Further data clustering is carried out on the feature vector using max pooling layer of the deep learning architecture on determining the inter-cluster seperability and intra-cluster compactness of the data vector in form of features. Moreover deep learning of the feature space in form of clusters is processed with gradient descent. Finally tuning of feature vector has been carried out on basis of using hyper parameter optimization with fewer epochs to determine the Discriminant information. Resultant representation learning and soft clustering achieved using deep adaptive fuzzy clustering is suitable to any kind of data distributions [16].

## 3. Proposed Model

This section provides an informal definition of the distributed deep conceptual learning based data clustering approach and later presents the deep concept specific learning framework for mining evolving data streams

### 3.1.Data Pre-processing

A large variety of datasets of form of high dimensional data are curated. Data Pre-processing has been applied in form missing value prediction and dimensionality reduction to determine effective clusters.

- **Missing Value Imputation**

Missing Value Imputation has been used factor analysis. Factor Analysis determines maximum common variance on the particular data field. It follows the Kaiser criterion which uses the Eigen value. It uses the score for the variance of the particular data field to fill the missed value of the data field. It can also compute using maximum likelihood method on basis of correlation of the data field [10].

- **Dimensionality Reduction**

Dimensionality reduction technique uses principle component analysis (PCA) to reduce the high dimensional data to lower dimensional data .It further used to eliminate over fitting. PCA is linear transformation technique. It reduces the feature based on correlation. It aims to project the subspace with fewer dimensions in high dimensional data. It has been processed using dimensional transformation matrix [11]. The domain-adapted parameters of the feature learning [10] are used as the model initialization for streaming data and the output feature vector F is clustered.

### 3.2. Feature Selection

Feature selection of the deep learning uses multiple layers to extract the features. Features extracted from hidden layer and deepest layer[12]. It can lead to better feature representations that can enhance the separation of data points during the similarity computation. It learns hidden features to encode and decode the data without considering the probability distribution of the input samples. Hyper Parameter value for fully connected layer of architecture computes the word vector of text data.

### 3.3. Deep Conceptual Incremental Clustering

Deep Conceptual Incremental Clustering has been employed for clustering the feature extracted in hidden layer. In this part, Auto encoder based deep learning model has been used. Initially encoder function has been employed to maps the extracted features into concept specific features representation and these representations have been processed in decoder function to reconstruct features into clusters [13]. Encoder and decoder function has been constructed as fully connected neural network.

**Table 1: Hyper parameter value for the fully connected layer**

| **Hyper Parameter** | **Values** |
|---|---|
| Batch Size of the cluster | 128 |
| Learning Rate of model | 0.01 |
| Size of word vector | 100 |
| Number of Epoch in max layer | 100 |
| Maximum Number of words in vector | 20000 |
| Maximum Sequence length | 1000 |
| Loss function | Cross entrophy |

In fully connected layer, Feature vector has been processed with hyper parameter values to generate the appropriate cluster for the selected word vector and to each epoch. Further cross entropy loss function has been utilized to manage cluster seperability of distributed data. Model parameter of neural network is updated to generate the distributed cluster with minimum inter cluster distance for novel data points. The optimization objective function composed of hyper parameter used in the fully connected neural network is given by

$$C = \lambda Lc + (1 - \lambda)Lc$$

Where $\lambda$ is considered as hyper parameter and $L_c$ is considered as Cluster limit

This function encourages the feature points on representative map to form cluster or become more discriminative in the particular cluster limit. Figure 1 represents the architecture of the proposed model.

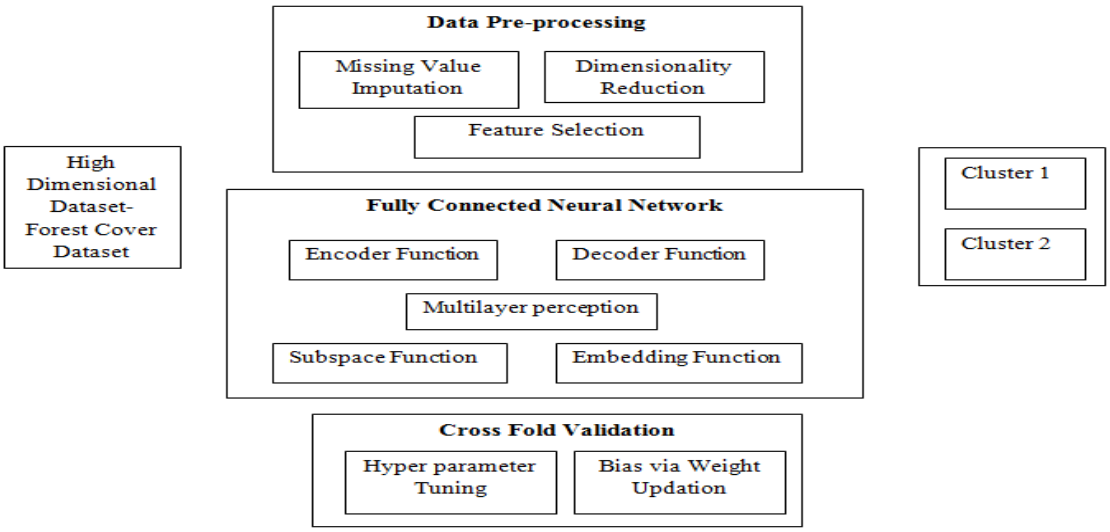


**Figure 1: Architecture diagram of the proposed deep concept specific learning**

Variants of autoencoder on different perspective have been used for optimization of hyperparameter on the objective function through bias and weight function. Auto encoder uses the word embedding as embedding function of neural network to find the concept of the feature to map into the cluster. Further it learns the non linear dependency of the data points and uses the transfer learning on encoder. Finally it eliminates the data sparsity issues effectives on weight updating through bias function [14].

Especially transfer learning uses the three constraints; Initial constraint utilizes a deep autoencoder to learn reduced representation from the streaming data on the feature selected instance as subspaces. Second constraints is applied in order to preserve the local structure data point of the vector from the original data, a locality preserving constraint is applied on the subspaces[15]. Third constraint uses the affinity points on subspace model to determine the subspace sparsity and affinity of representations.

**Algorithm 1: Deep Concept Specific Learning**

Input : High Dimensional Dataset
Output : Data Clusters
Process
  Data Pre- Process ()
    Compute Missing value ()
        Assign Kaiser Criteria
            Set Eigen value of the variance as Data input to missing field
    Select Feature()
      Feature Reduction _PCA ()
        Dimensionality Reduction o
      Feature extract_PCA()
Return feature F
 Apply Deep Concept Specific Learning ()
   Generate subspace for F
     Calculate Latent F on Hidden Feature on Subspace
       Transfer learning ()
              Encode ()
                 Map Latent F on word embedding

Decode ()
Bias the Map
Compute distance of instance and group //Cluster based on similarity
Return Cluster

The algorithm of deep specific learning has been applied to generate the cluster which maximizes the accuracy and minimizes the reconstruction error. However, since covering each of them in hyper parameter tuning would be cumbersome in this comparative analysis, Hence, detail of bias and weight updates as training for the AE methods has been detailed and Cross validation on test data includes most of the possible steps explained in proposed approaches

## 4. Experimental Results

Experimental analysis of deep learning architecture for cluster generation has been carried out on the forest cover data which is high dimensional in nature. The performance of the proposed technique has been evaluated utilizing precision, recall and Fmeasure. The proposed model is experimented and evaluated using Dotnet technology. In this particular platform, processing of the data clustering in form of text is highly challenging to train and validate the system.

In this work, 60% of input dataset has employed to train corresponding learning architecture to recognize the text in word embedding process and clustering and 20% of dataset is used to validate the proposed model on cross fold validation. Finally 20% of the data is used for testing. In this work, 80% is considered for training data has divided into 60% to train the model and 20% to validate the trained model. Finally performance of the model is cross evaluated using 10 fold validation. Figure 2 represents the performance evaluation of the proposed architecture in terms of precision on forest cover dataset. The training parameter of the reprehensive learning has been defined in the table 2

**Table 2: Training parameters**

| **Parameter** | **Value** |
|---|---|
| Learning rate of the data | $10^{-6}$ |
| Loss Function | Categorical cross entropy |
| Batch size of vector | 15 |
| Max epoch for fully connected architecture | 1000 |

### 4.1. Dataset Description

We have carried out extensive experiments on forest cover datasets in order to measure the outcome of the clustering performance. In this model, each dataset produces the data segmented into equal parts for training and testing. In this experiment, training of model consumes 60%, Validation consumes 20% and testing consumes 20%. Detailed description of the dataset is represented as follows

- **RCV1 (Reuters Corpus Volume I).**

Reuter data set contains data corpus of newswire which describing the collection of the news feed which is mostly frequently used bench mark dataset for many data clustering techniques.

- **Forest covers data set from UCI repository (Forest).**

The data set contains geospatial descriptions of different types of forests cover information's. We normalize the data set, and arrange the data so that new classes appear randomly on basis of probability distribution [14].

- **Twitter**

Twitter data set contains 340,000 Twitter messages (tweets) of different trends on wide classes in different area and subjects on world.

### 4.2. Evaluation

The proposed technique has been evaluated against the following performance measures against traditional deep representative learning. In this work, proposed model is evaluated using 10 fold validation to compute the performance of cluster generated on various dataset mentioned. The performance evaluation of the proposed deep learning model depends on the process of activation function, pooling layer, weight function, bias function and hyper parameter of model. In addition fuzzy membership has been used as criteria to partition the cluster on fully connected structure.

- **Precision**

It is a measure of Positive predictive value. It is further represented as the fraction of relevant instances among the each cluster groups generated using the model. In another way of presentation, Precision is considered as number of correct feature divided by the number of all returned feature space obtained. Figure 2 represents the performance evaluation of the proposed architecture in terms of precision on forest cover dataset. Performance measures are suitable for determining the feasibility of proposed architecture on cluster generation. Effectiveness is achieved due to hyper parameter tuning

$$\text{Precision} = \frac{\text{True positive}}{True\ positive + False\ Positive}$$

True positive is a number of similar points in the data and false negative is number of real dissimilar points in the data[15]. Mostly a good clustering performance is also characterized by high intra-cluster similarity and low inter-cluster similarity for the data points. It can be calculated using recall measure.

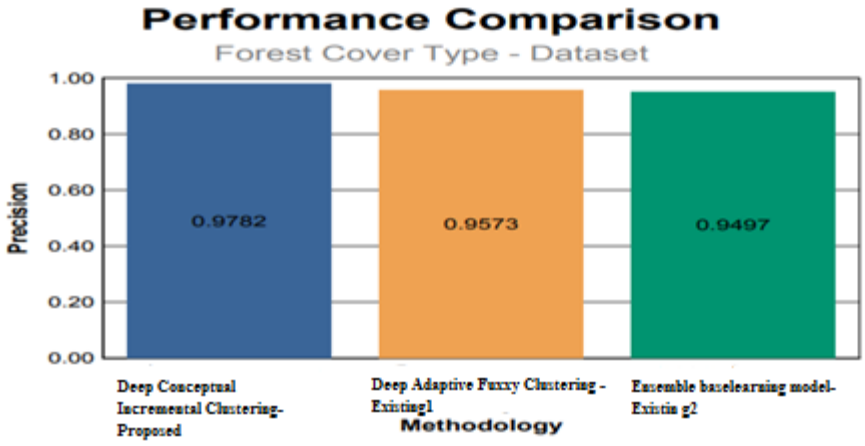


**Figure 2: Performance analysis of the methodology on aspect of Precision**

.

- **Recall**

Recall is the part of relevant data points that have been extracted over the total amount of relevant data point of cluster. The recall is the part of the relevant documents that are successfully classified into the exact classes.

$$\text{Recall} = \frac{\text{True positive}}{True\ positive + False\ negative}$$

True positive is a number of similar data points in the data and false negative is number of similar data points in the data. Figure 3 represents the performance evaluation of the proposed architecture on recall measure along state of art approaches.

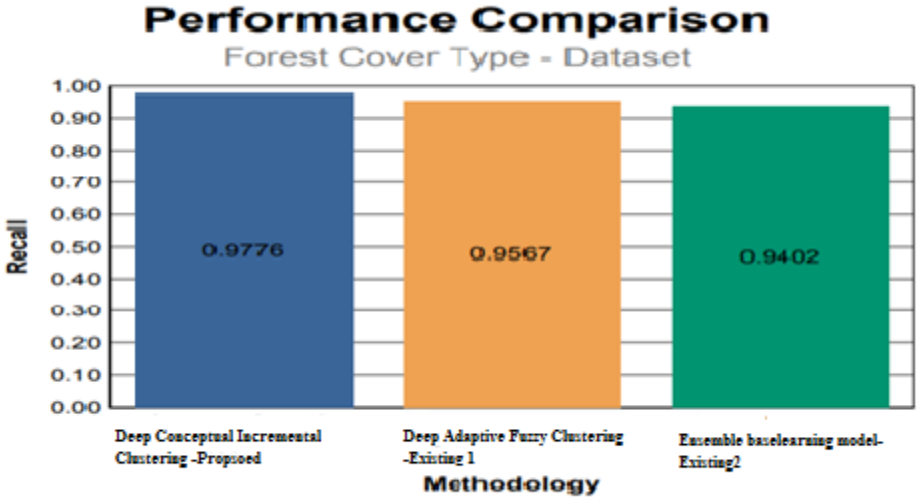


**Figure 3: Performance analysis of the methodology on aspect of Recall**

Cluster quality depends on activation function in every layer. Encoder calculates the feature map to generate subspace. F measure is a good measure for determining the quality of the clustering.

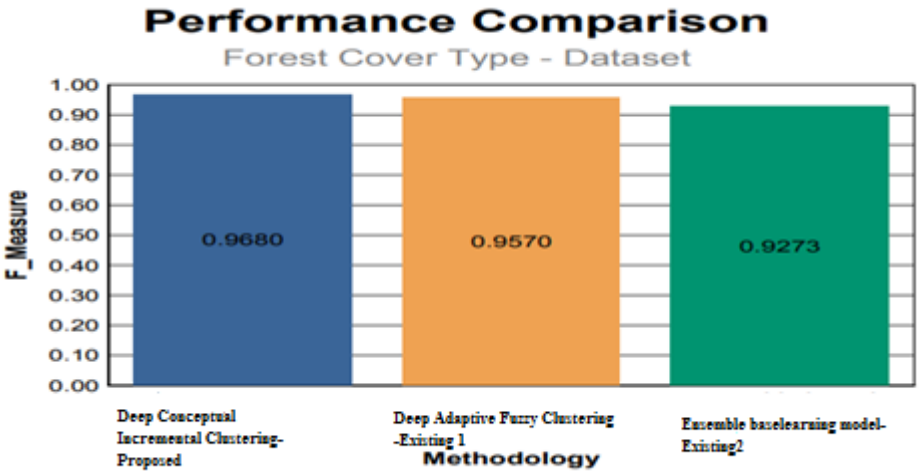


**Figure 4: Performance analysis of the methodology on aspect of F- Measure**

- **F measure**

It is the number of correct class predictions to the incoming data among total number of predictions to whole category of data.

Accuracy is given by

$$\frac{\text{True positive} + \text{True Negative}}{True\ positive + True\ Negative + false\ positive + False\ negative}$$

Although different data points may have different impact on cluster formation, they are likely to have the same impact on clustering. Figure 4 represents the performance of the proposed model in terms of f measure against state of art approaches for high dimensional data clustering. However, after a certain point, this data vector is diminished because of curse of dimensionality. On the other hand, for the forest data set, the clusters generated are less separable.

Proposed encoder function acts as an approximation function to map the input into a distribution. Then, the generative probabilistic decoder tries to generate the original sample by means of conditional probability. Table 2 presents the performance value of the technique for cluster analysis.

**Table 2: Performance Analysis of Deep learning architecture against state of art approaches**

| Technique | Precision | Recall | F measure | Accuracy |
|---|---|---|---|---|
| Deep Conceptual Incremental Clustering model - Proposed | 0.9782 | 0.9776 | 0.9680 | 0.9921 |
| Deep Adaptive Fuzzy Clustering –Existing 1 | 0.9573 | 0.9567 | 0.9570 | 0.9847 |
| Ensemble based Learning Model- Existing 2 | 0.9497 | 0.9402 | 0.9273 | 0.9448 |

On the other hand, the clustering method not only for detecting clusters in a given high dimensional data, but it is capable of detecting the underlying structure of the data distribution in general. It is naturally much more powerful, since they can handle nonhyperspherical clusters.

## Conclusion

Deep Concept Specific Clustering is deep learning technique designed and implemented for analysing unstructured and high dimensional data into cluster. Proposed model uses the auto encoder based fully connected neural network via generating the cluster with minimized reconstruction error. Model uses the word embedding for concept specific feature generation towards mapping the sparse representation into clusters. Cluster performance computed using f measure proves that it is effective on cluster instance similarity by better initialization of the parameters. Finally proposed model proves that it is effective and high scalable on high dimensional data.

## References


1. Min E, Guo X, Qiang "A survey of clustering with deep learning: from the perspective of network architecture" in IEEE Access, Vol. 6, issue.39, pp: 501–14, 2018.
2. Chowdary NS, Prasanna DS, Sudhakar P. "Evaluating and analyzing clusters in data mining using different algorithms". International Journal of Computer Science and Mobile Computing, Vol.3, PP: 86–99, 2014.
3. Davidson I, Ravi SS. Agglomerative hierarchical clustering with constraints: theoretical and empirical results. In: European Conference on Principles of Data Mining and Knowledge Discovery. Heidelberg, Germany: Springer, 2005, pp: 59–70.
4. Dizaji KG, Herandi A, Cheng," Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization. In: 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017, 5747–56.
5. Xie J, Girshick R, Farhadi A "Unsupervised deep embedding for clustering analysis. In: International Conference on Machine Learning. New York City, NY, USA: ICMLR, 2016, 478–87.
6. K. Buza, A. Nanopoulos, and L. Schmidt-Thieme, "INSIGHT: Efficient and Effective Instance Selection for Time-Series Classification," Proc. 15th Pacific-Asia Conf. Knowledge Discovery and Data Mining (PAKDD), Part II, pp. 149-160, 2011.
7. L. Parsons, E. Haque, and H. Liu, "Subspace clustering for high dimensional data: a review," ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 90–105, 2004
8. N. Tomasev, M. Radovanovic, D. Mladenic, and M. Ivanovic, "The role of hubness in clustering high-dimensional data," IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 3, pp. 739–751, 2014.
9. H. Kremer, P. Kranen, T. Jansen, T. Seidl, A. Bifet, G. Holmes, and B. Pfahringer, "An effective evaluation measure for clustering on evolving data streams," in Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2011, pp. 868–876.
10. P. Huang,Y. Huang,W.Wang, and L.Wang, ``Deep embedding network for clustering," in Proc. 22nd International. Conference. Pattern Recognition. (ICPR), Aug. 2014, pp. 1532-1537.
11. P. Ji, T. Zhang, H. Li, M. Salzmann, and I. Reid, ``Deep subspace clustering networks," in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 23-32.
12. W. Harchaoui, P. A. Mattei, and C. Bouveyron, ``Deep adversarial Gaussian mixture auto-encoder for clustering," in Proc. ICLR, 2017, pp. 1-5.
13. N. Dilokthanakul et al. (2016). ``Deep unsupervised clustering with Gaussian mixture variational autoencoders." [Online]. Available: https:// arxiv.org/abs/1611.02648
14. G. Chen. (2015). ``Deep learning with nonparametric clustering." [Online]. Available: https://arxiv.org/abs/1501.03084
15. Z. Jiang, Y. Zheng, H. Tan, B. Tang, and H. Zhou. (2016). ``Variational deep embedding: An unsupervised and generative approach to clustering." [Online]. Available: https://arxiv.org/abs/1611.05148
16. S.Praveen & R.Priya "A unified deep learning framework for text data mining using deep adaptive fuzzy clustering" in Solid state technology, Vol.63, issue.6, 2020.