Nat. Volatiles & Essent. Oils, 2021; 8(5): 1695 - 1703

NVEO
Natural Volatiles & Essential Oils

High Dimensional Deep Data Clustering Architecture Towards Evolving Concept

A. Sumathi [*1], K. Yasotha [*2], S. Nandhinidevi [*3]

*[*1]Associate Professor, Department of Computer Science, Dr. N.G.P Arts and Science College, Coimbatore, Tamilnadu. Email: sumathi@drngpasc.ac.in*

*[*2]Assistant Professor, School of Computer Science, PPG College of Arts and Science, Coimbatore, Tamilnadu. Email: yasothak.cas@ppg.edu.in*

*[*3]Assistant Professor (SRG), Department of Computer Technology - UG, Kongu Engineering College, Perundurai, Tamilnadu. Email: nandhinidevi.ctug@kongu.edu*

Abstract

Data Clustering on the evolving data stream has becoming primary and vital tasks in the natural language processing in data driven application domains. Performance analysis of data clustering techniques depends on the quality of data representation on the evolving data streams. Machine learning algorithm plays a primary role in high dimensional data clustering of the evolving data streams from social media. Despite of large benefits on implementing those algorithms, it suffers on various aspects such as low accuracy and distribution of the data points in the time varying clusters. In order to resolve above mentioned issues, Deep learning architectures has been analysed on various aspects to develop a optimized model for better data representation of data clustering. In this paper, an optimized deep learning architecture which termed as Concept Based High Dimensional Deep Clustering has been projected for analysing unstructured and high dimensional evolving data in the data streams. Proposed architectures use the hidden layer for better effective identification of the hidden feature. Further the evolving data uses transform learning for data transformation. Those learning transforms high dimensional to low dimensional deep feature space. Finally these deep feature spaces extract the concept specific features on the employment of the max layer using principle component analysis (PCA). PCA on the max layer determines the salient features on ensuring the minimum reconstruction error. Deep architecture is eliminating the NP hard problem and over fitting issues of the clustered results. Further all parameters are fine tuned with respect to certain criterion on cross validation. The Softmax layer is used to map the data points into accurate cluster 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 proposed deep clustering model 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

Data Clustering is a fundamental data learning architectures employed for grouping the evolving data stream as data mining applications. The primary objective of data clustering is to categorize distributed data points into one or more cluster based on distance similarity measures on the data points (e.g., Euclidean distance, Cosine Similarity etc). However a large number of data clustering methods using machine learning architecture have been presented [1]. Traditional data clustering methods using machine learning approaches such as K means and Fuzzy C means will produce poor performance on high-dimensional data [2], as it is inefficient due to distance computation of resultant clusters on similarity measures used in these methods. Furthermore, these machine learning methods generally faces the problems such as curse of dimensionality, data sparsity and computational complexity on large-scale datasets [3][4]. To tackle those implications, dimensionality reduction and feature transformation methods [5] like Principal component analysis

(PCA) [6] and Linear Discriminant Analysis [7] and spectral methods [8] have been extensively studied to cluster the high dimensional data into a new feature representation. However, a highly complex feature representation of data extracted is highly challenging on employment of the machine learning data clustering techniques. On development of deep learning architectures, learning representation of the evolving streams can be used to transform the data cluster into cluster friendly data distributions.

Deep learning architecture for evolving data clustering has been employed using artificial neural network for better cluster representation of high dimensional evolving data. In this article, high dimensional deep data clustering has been proposed as deep learning approach for analysing unstructured and high dimensional evolving data. Approach uses concept specific learning category to obtain the feasible feature space containing concept specific features. The Deep representative model provides non linear mapping function to construct the features with loss criteria to include the latent representation. Eventually, learning deep evolving data streams has been clustered on the feature transforming constraints employed on input data to produce more cluster-friendly representations in which the evolving data on representing into a lower-dimensional feature representation

The Remaining of the article is sectioned as follows, related work of the data clustering techniques are described in section 2, the architecture of the proposed High Dimensional Deep data clustering approach is described in section 3 and experimental results and effectiveness of the proposed approaches is evaluated against the state of art approaches in section 4 using evolving social media dataset. Further performance comparison of the proposed model has been carried out on various performance measures with state of arts approaches has been explained. Finally this research article is concluded in section 5.

2. Related Work

In this section, High Dimensional Data clustering model using machine learning approaches has been analysed against various processing details on basis of architectures for feature extraction and distance calculation using similarity measures of the clustered data points.

2.1. Ensemble based data Clustering

In this literature, Ensemble based data clustering prototype belongs to machine learning paradigm has been described for high dimensional data clustering. It partitions the input data points into a multidimensional space to form 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 fast learning framework for Imbalanced data

In this literature, evolving data stream will be processed to extract the feature using ensemble feature extraction technique such as Incremental Kernel Principle Component Analysis [5], Incremental linear Discriminant analysis and Incremental Linear Principle Component Analysis [6]. Feature subset extracted undergoes ensemble classification through chunk based ensemble

classifier and online ensemble classifier. This classifier form uses recurrent neural network, deep belief network, convolution neural network and autoencoder. Base classifier and class imbalance has been handled using weighted average and under sampling methods on replacing the older model with newly trained model. It is also capable of handling multiclass drift.

3. Proposed Model

This section provides an informal definition of the evolving distributed high dimensional data clustering approach and later presents the deep 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.

- **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) in order 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].

3.2. Data Dimensional Feature Selection

High Dimensional Feature selection representatives extract the features of the evolving data. Those extracted Features using deep learning architectures provides sparse representation [12]. Feature obtained can enhance the separation of data points on clustering and it can be evaluated during the similarity computation. Feature describes the greatest amount of variance on the eigen value of eigen vector. The variance computation of the features is given by

$$\text{var(x)} = \frac{\sum_{i=1}^{n} \text{a(xi - x) (xi - x)}}{n-1}$$

On basis of variance computation, feature vector has been generated and it is aggregated using sliding window method on the various feature spaceIt learns latent hidden features to categorize the data without considering the evolving probability distribution of the input data points.

3.3. High Dimensional Data Clustering

High Dimensional Data Clustering has been employed for clustering the feature extracted in hidden layer. In this part, deep learning model has been used to cluster the data point of evolving data streams. Initially clustering function has been employed to categorize the extracted features into sparse features representation and these representations have been processed in decoder function to reconstruct features into clusters [13]. Deep clustering function has been constructed as fully connected artificial neural network. 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 λ is considered as hyper parameter and L_c is considered as Cluster limit

This function encourages the feature points on representative strcuture to form cluster or become more discriminative in the particular cluster limit. In base layer of the learning, several classifiers are generated to selected feature set for classification to establish an ensemble Figure 1 represents the architecture of the proposed model.

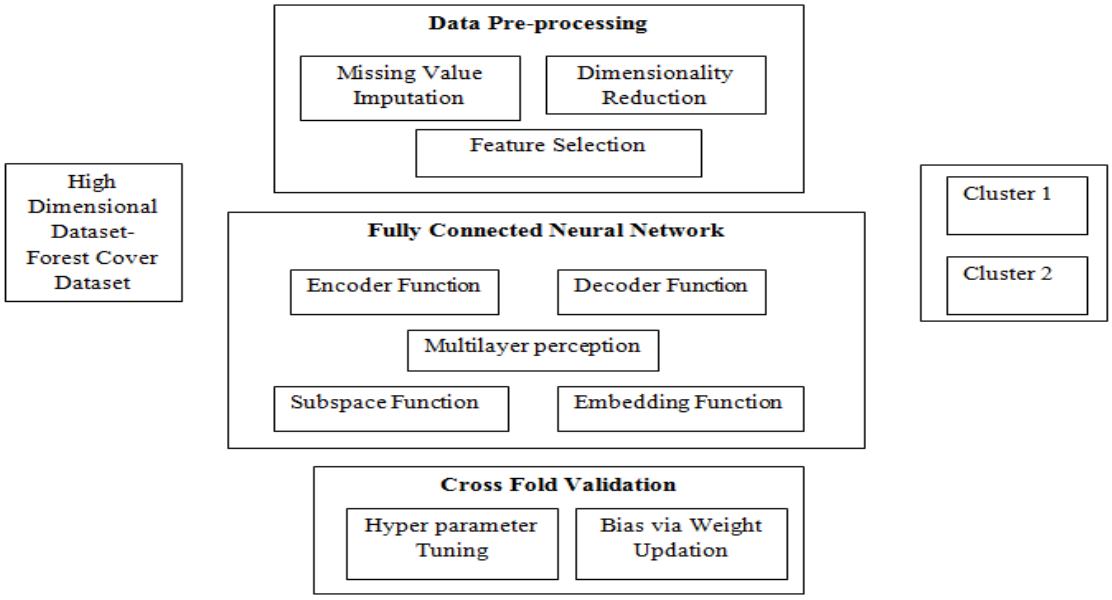

Figure 1: Architecture diagram of the proposed deep concept specific learning

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

Especially transfer learning for cluster uses the three constraints; it first utilizes a deep sparse clustering constraint to learn reduced feature representation from the raw data on the feature selected instance as subspaces. Second constraint of the clustering is to preserve the local feature structural property of the original evolving data which considered as the subspaces [15]. Each roll out weights the remaining nodes in the tree which consist of feature value to easily categorize it to classes either generating new classes and on existing classes. Finally, it incorporates a feature subspace sparsity constraint to determine the data affinity of cluster representations.

Algorithm 1: High Dimensional Deep Data clustering

```
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 High Dimensional Deep Data Learning ()
   Generate subspace for F
     Calculate Latent F on Hidden Feature on Subspace
       Transfer learning ()
                Cluster Constraint ()
                   Represent Latent F on word embedding
                Subspace Constraint ()
                      Extend the Feature on biasing
                      Compute distance of instance and group //Cluster based on similarity
      Return Cluster
```

The algorithm named High Dimensional Deep Data Clustering architecture has been applied to generate the cluster which maximizes the accuracy and minimizes the reconstruction error. However, data cluster has been optimized on hyper parameter tuning to generate the sparse cluster for evolving data streams. Finally, details of bias and weight updates as training approaches for the deep learning methods has been detailed for effective cluster generation and Cross validation on test data includes most of the possible steps explained in proposed approaches to compute the effectiveness. It is considered as ways of biasing the feature that lets the class tree expansion towards the most promising features.

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. Further 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.

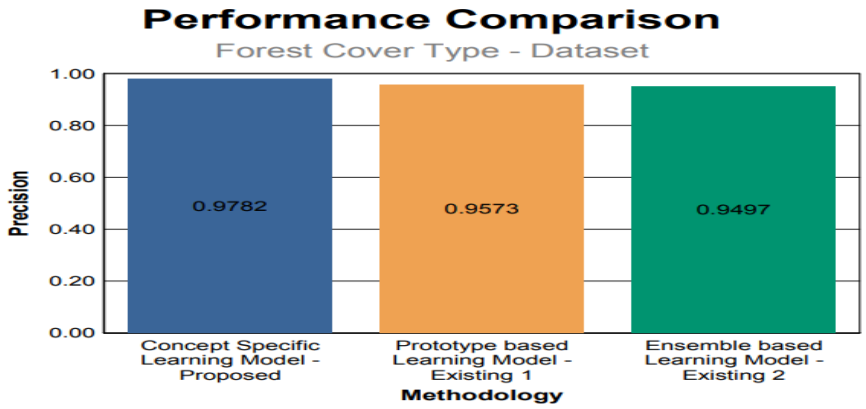

Figure 2: Performance analysis of the methodology on aspect of Precision

Performance measures are suitable for determining the feasibility of proposed architecture on cluster generation. Effectiveness is achieved due to hyper parameter tuning.

- **Precision**

The precision is a measure of positive predictive value of feature class. It is also considered as ratio of retrieved instance to resultant instance of the feature vector towards classification [15]. The performance outcome is represented in the figure 2

Precision = P= $\frac{Releveant\ retrieved\ result\ set}{Overall\ Retrived\ Result\ set}$

Mostly a good clustering performance is also characterized by high intra-cluster similarity and low inter-cluster similarity measures for the clustered data points. It can be calculated using recall measure. Figure 3 represents the performance evaluation of the proposed architecture on recall measure along state of art approaches.

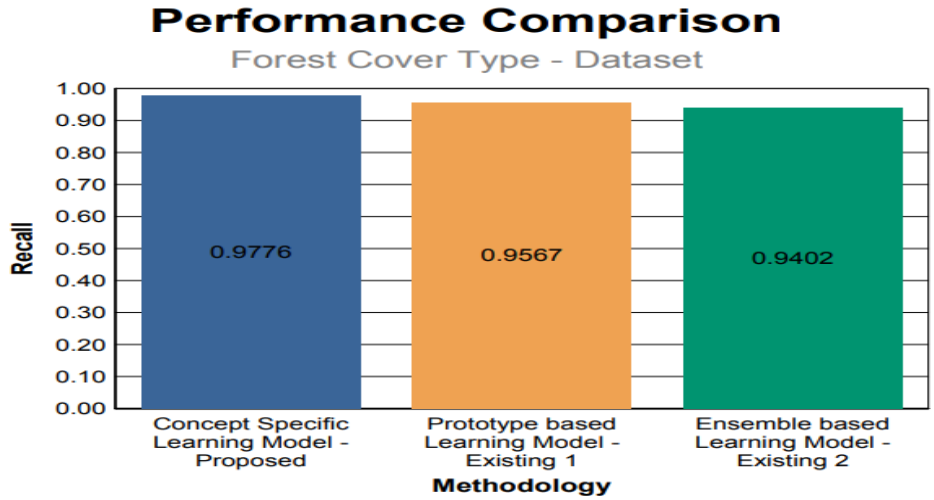

Figure 3: Performance analysis of the methodology on aspect of Recall

Cluster quality depends on activation function in every layer of the deep architecture. A deep Cluster result calculates the feature parttions to generate subspace.

- **Recall**

The Recall is ratio of relevant feature instance extracted in the vector to retrieved feature vector on class of instance. Its performance outcome is represented in figure 3

Recall = R= $\frac{Releveant\ retrieved\ result\ set}{Relevant\ Result\ in\ database}$

F measure is a effective measure for identifying the quality of the high dimensional data clustering. Figure 3 represents the performance of the proposed model in terms of f measure against state of art approaches for high dimensional data clustering.

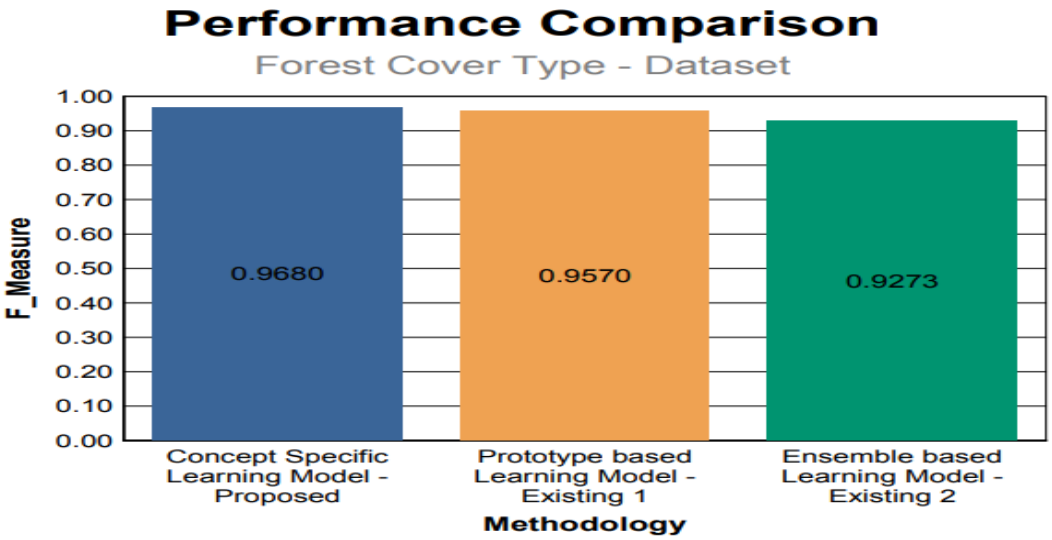

Figure 4: Performance analysis of the methodology on aspect of F- Measure

Proposed Deep Subspace function acts as a Cluster approximation function to partition the input data into a distributed data cluster. Then, the generative probabilistic feature tries to generate the original cluster by means of conditional probability of the sparse representation [16, 17]. Table 1 presents the performance value of the technique for deep cluster analysis.

Table 1: Performance Analysis of Deep learning architecture against state of art approaches

S. No	**Technique**	**Precision**	**Recall**	**F measure**	**Accuracy**
1	High Dimensional Deep Data clustering Approach l - Proposed	0.9782	0.9776	0.9680	0.9921
2	Prototype based Data Clustering	0.967	0.9677	0.9584	0.9856
3	Ensemble based Learning Model-Existing 2	0.9497	0.9402	0.9273	0.9448

On the other hand, the deep data clustering methods not only for identifying cluster groups in a given high dimensional evolving data, but rather it is effective predicting the underlying sparse structure of the data cluster distribution in general.

5. Conclusion

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

References

1. Min E, Guo X, Qiang, et al. A survey of clustering with deep learning: from the perspective of network architecture. IEEE Access 2018; 6:39501–14.
2. Chowdary NS, Prasanna DS, Sudhakar P. Evaluating and analyzing clusters in datamining using different algorithms. Int J Comput Sci Mob Comput 2014; 3:86–99.
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, 59–70.
4. Dizaji KG, Herandi A, Cheng, et al. 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 Int. Conf. Pattern Recognit. (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. E. Boopathi Kumar, V. Thiagarasu, “Segmentation using Fuzzy Logic in Color Images Based on Membership Functions”, International Journal of Engineering Sciences & Research Technology, pp. 38 - 45, Volume 6, Issue 6, 2017.
17. E. Boopathi Kumar, V. Thiagarasu, “Comparison and Evaluation of Edge Detection using Fuzzy Membership Functions”, International Journal on Future Revolution in Computer Science & Communication Engineering, pp. 149 – 153, Volume 3, Issue 8, 2017.
18. E. Boopathi Kumar and M. Sundaresan, “Edge Detection Using Trapezoidal Membership Function Based on Fuzzy‘s Mamdani Inference System”, 2014 International Conference on Computing for Sustainable Global Development (INDIACom), pp. 515-518.