Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.778038

A MULTIVIEW CONVOLUTION NEURAL NETWORK FOR INCREMENTAL STREAMING DATA CLUSTERING

R. Ramesh
Associate Professor, Department of computer science, KPR College of Arts Science and Research, Coimbatore, proframeshr@gmail.com
S. Sridevi
Assistant Professor, Department of Computer Science, KPR College of Arts Science and Research, Coimbatore, devisris.mphil@gmail.com
R Rajkumar
Assistant professor, Sri Krishna arts and science college Coimbatore,
K. Dhiyaneshwaran
Dept. of Computer Technology and Information Technology, Kongu Arts And Science College (Autonomous), Erode 638107, dhiyaneshwarank@kasc.ac.in

Abstract

Incremental streaming data clustering is an important research consideration in social media applications and performance of clustering on the streaming data is largely depends on the data representation quality with respect to clustering effectiveness and data efficiency. Machine learning technique is a employed as traditional approach to streaming data clustering but faces huge complications due to increasing evolution of data contents. In order to tackle those issues, a multi-view convolution neural network has been presented in this paper. The Proposed model uses convolution layer for feature reduction and extraction. Pooling layer for feature selection which select the feature with maximum weight. Selected feature is projected to fully connected layer. Fully connected layer maps the feature into generated clusters on basis of objective function with maximum margin cluster. Those cluster further fine tuned to modify the hyper parameters of various layers in the convolution neural network to maintain the classification error on managing the ReLu activation layer in specified limit. Softmax layer minimizes the feature variance and cluster feature seperability in the feature space. Hyper parametric tuning is carried out in the output layer to make the data instance in the cluster to be close to each other by determining the similarity of the data instances on cluster representation. It results significantly enhancement in the clustering performance using the discriminative information's. Detailed experiments of the current model have been evaluated against state of art techniques using facebook datasets. The implementation outcome of the multiview convolution neural network for data learning architecture represent the high accuracy and efficiency on clustering the streaming data.

Keywords: Deep Learning, Streaming data, Deep Clustering, Convolution Neural Network Unstructured Data

1. Introduction

Clustering is a significant data management approach to acquire significant information on the data distributions by enabling to cluster or group a large volume of data into highly representative clusters. Clustering of streaming data is considered to be sparse and it

contains large number of data features for cluster representation. It explores exponentially with the increased no of data. Further it become challenging to cluster the feature point

Approach to manage the streaming data is ubiquitous and overloading [1]. Machine learning model employs the feature extraction [2] to extract the features for data clustering but it leads to over fitting issues. Towards large exploration of the features for clustering task, deep learning model has been employed. Deep learning mechanism for streaming data clustering is explored towards achieving soft clustering by generating a new cluster representation [3].

In this article, multi-view convolution neural network architecture has been generated as a deep learning architecture for analyzing non-uniform streaming text data. Proposed model employs convolution neural network to map the data into embedding objects as feasible feature space containing latent features on max pooling layer. The CNN model contributes non linear objective function using ReLu to establish the data features with reconstruction error criteria have to produce maximum margin cluster on refining parameters.

The remaining part of the article is represented as follows, analysis of similar literature are discussed in section 2, the current approach considered as convolution neural network architecture is represented in section 3. Implementation outcomes and performance outcome of the current architecture is represented in section 4 with respect to accuracy measures. Finally, paper is summarized in section 5.

2. Related works

In this segment, streaming data clustering approaches using machine learning architectures has been analyzed in detail. Those technique is represented as follows

2.1. K Means Clustering

In this approach, K means clustering is analyzed for streaming data clustering. k means model provide the cluster partition containing data points. Initially kernel method has been utilized to extract the implicitly features to construct the feature space [5]. Feature space is reconstructed as hierarchical random feature set. On employing clustering rules, best feature representation into cluster has been obtained.

3. Proposed Architecture

Proposed architecture models a detailed specification of the proposed streaming data clustering architecture on incorporation of hyperparameter tuning of the deep learning layers to generate the soft partition clusters.

3.1.Data Pre-processing

Incremental datasets in structure of streaming data are complex in data clustering. Data Pre-processing is implemented using KNN to obtain the effective clusters data on removing stop words and stemming process.

3.2. Multi-View Convolution Neural Network

CNN is deep learning architecture implemented for clustering the feature. In this part, Convolution Neural Network uses hyper parameter tuning using epoch and activation function to generate cluster. Generated objective functions based deep learning architecture is been utilized for feature selection using convolution layer. In beginning, max pooling is included as first layer to generate the feature map to the extracted features from the convolution layer and these vectors have been managed in convolution layer to down sample the features [13].

- **Max pooling layer**

Feature Subset containing variance probe undergoes various constraints has been proposed in the max pooling layer of the neural network to generate sparse feature for convolution layer. The feature depiction which can increase the variation of data points on the similarity calculation of data clustering have been provided as outcome. The hyperparameter components of the convolution Neural Network for cluster tuning.

Table 1: Hyper Parameterized for the Multi-View Convolution Neural Network

S. No	Hyper Parameter	Values
1	Cluster Size	15
2	Learning Rate	0.001
3	Number of Epoch	45
4	No of data points in cluster	10000
5	Loss function	Cross entropy

- **Convolution Layer**

The Convolution layers gathers the feature using its inherent process hierarchically among low levels to large abstract features in order to compute the large variance features with very few parameters. It is represented in form of tensor. It operates in a condition that it generates weight around each variable of steaming feature space. Result of the feature vector is feature instance of the cluster.

- **Batch Normalization**

Batch normalization is implemented towards quick convergence of the feature to clustering for the feature from convolution and max pooling layer using epoch value. Feature normalization is the fixing feature vector in appropriate range on utilizing the ReLU activation function. It eliminates the feature instances.

- **Activation Function**

Current architecture implements the rectified linear units (ReLU) activation function, which produces the non-linearity to the extracted features. Feature vector is computed with non parameterized values to produce the perfect cluster to the generated feature vector on every epoch.

- **Output Layer**

The output layer of the convolution neural network contains the cluster with feature instances. Further hyper parametric tuning of the cluster based on distance measures is achieved in this layer. Soft max and cross entropy approach of loss function on the cluster instances has been processed to generate effective clusters.

- **Loss Layer.**

Loss layer is to guarantee the cluster result on fine tuning on the modifying parameter of various layers in convolution neural network to guarantee the reduced classification error in the middle of feature max pooling layer and ReLu activation layer. Further cross entropy loss function is been incorporated to control the cluster variance of feature vector.

Algorithm 1: Multi-View Convolution Neural Network Learning

Input: Streaming Dataset -Twitter
Output: Text Clusters
Process
Apply Deep Convolution Neural Network Learning ()
Max Pooling ()
Extract the features and map it as feature map
Convolution Layer()
Compute feature
Batch Normalization()
Normalize feature
Activation Layer()
Cluster using ReLu function
Loss Layer()
Non Parameterized Tunning of ReLu Function
Output Layer()
Softmax()---Cluster

This process motivates the feature instances on feature map to establish data cluster or to yield large differential features to the selected cluster data range. Figure 1 illustrate the architecture of the current approach.

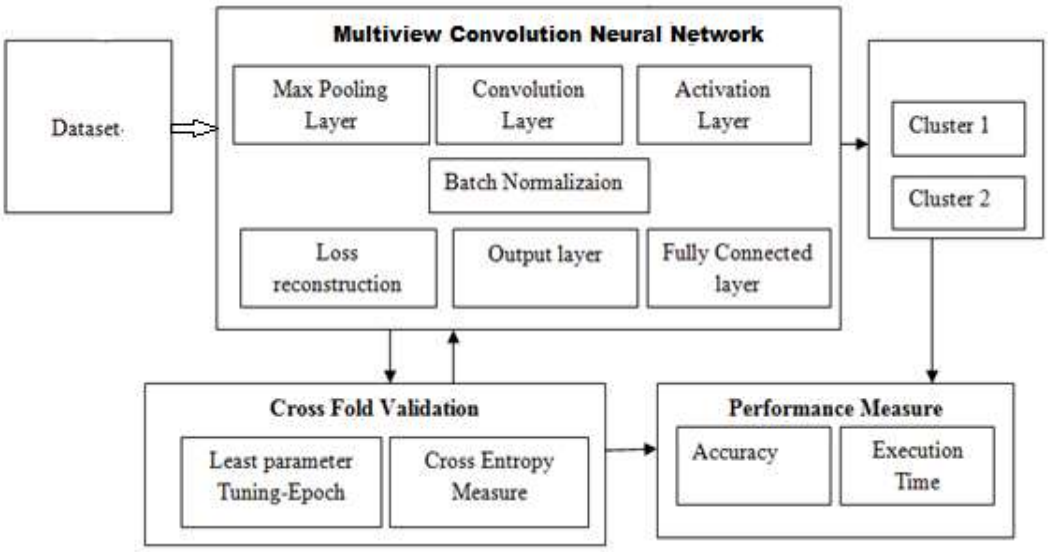

Figure 1: Architecture of the Multiview Convolution Neural Network

Training of the Deep learning achieves the clustering on the changes on streaming data in the architecture of the CNN, loss function and training approach on varied perspective details about the network update. Further it generates the feature discriminative dependency of and implements the transfer learning on activation function using batch normalization. Finally, it avoids the data deficiency complication effectively on enhancing the feature weight [14].

The CNN learning algorithm is implemented to produce the data cluster which enhances the cluster accuracy and minimize the classification error through loss function. However, since wrapping every feature in hyper parameter updating will be quite complex in this similarity analysis of the cluster results. As the training epoch increase, the error will steadily minimize and the model accuracy is enhanced [15].

4. Experimental Results

Experimental analysis of current multi-view convolution neural network learning approach using hyper parameter tuning on streaming data for cluster yield is projected with the twitter dataset which is streaming in representation. The performance of the current approach is computed including precision, recall and F measure against real time dataset. The current approach is modeled and computed its performance using python technology.

Finally model performance is evaluated using cross entropy function trough 5-fold validation. Figure 2 illustrates the performance assessment of the current approach on basis of precision using the twitter dataset. The hyper parameter of the current data clustering learning model is detailed in the table 2

Table 2: Training parameters

S. No	Parameter	Value
1	Learning rate	10^{-3}

2	Loss Function	cross entropy
3	Vector size	100
4	epoch	50

4.1. Dataset Description

The detailed experiments are carried out using twitter datasets which assess the performance of clustering outcomes. In this architecture, every bench mark dataset yields the data partitions into similar portions for cluster training and cluster testing. In this implementation, training architecture utilizes 60% of the dataset, Validation utilizes 20% of the dataset and remaining portion of dataset is utilized for testing. Specific representation of the dataset is illustrated as follows

Twitter data set.

Particular data set represents geospatial descriptions of various classes of attributes information's. It is to normalize the dataset, and rank the instances according as new classes generates randomly with respect the probability distribution of the dataset [14].

4.2. Evaluation

The current approach is analyzed on basis of following performance measures against conventional k mean clustering learning architecture. In this article, current approach is accessed employing 5-fold computation to calculate the performance of cluster outcome on twitter dataset depicted. The performance assessment of the current deep learning architecture is based on the utilization of activation function, pooling layer, loss function and fully connected layer of model.

- **Precision**

It is a compute of Positive predictive score. It is illustrated as the ratio of similar data points in the cluster group's yield employing the proposed approach. Figure 2 illustrates the performance computation of the current approach with respect to precision using twitter dataset. Performance assessment are efficient for computing the reliability of current approach on data cluster yielded. Cluster effectiveness is obtained using the hyper parameter updating.

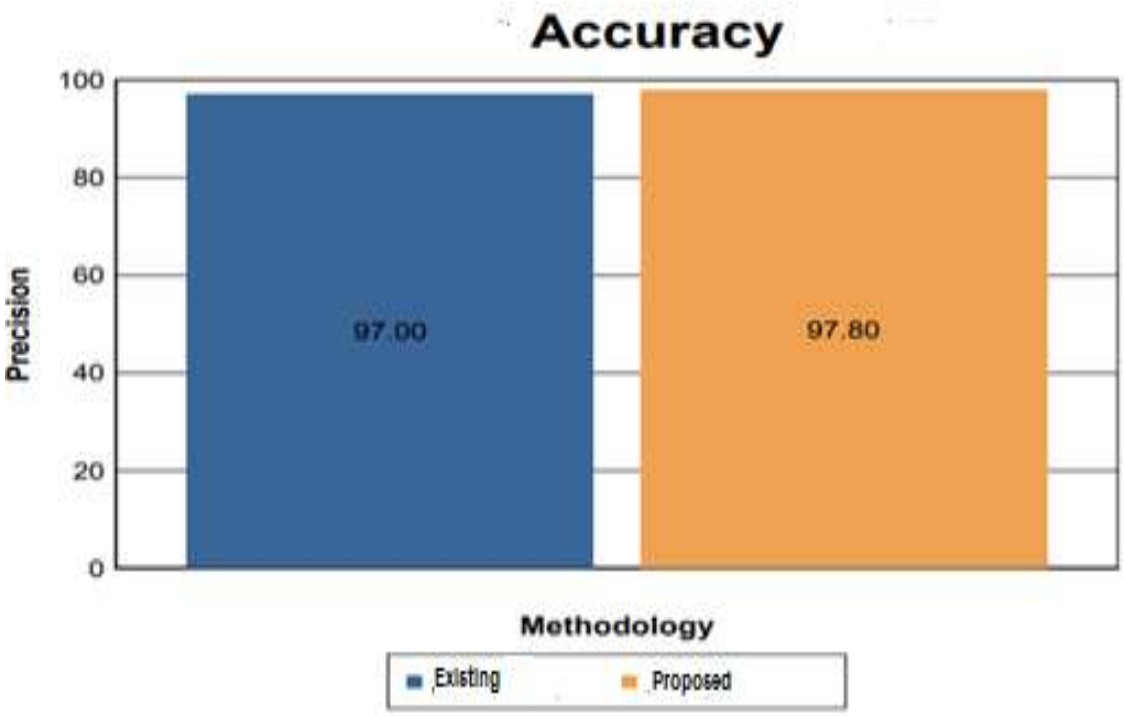

Figure 2: Performance analysis of the Precision Value

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

True positive is represented as number of relevant points in the dataset and false negative is number of not correctly related points in the dataset [15]. In particular, an effective clustering assessment is also represented by large accuracy for the cluster data points. It is computed utilizing the recall metrics.

.

- **Recall**

Recall is the represented as relevant data points which is extracted over the total volume of similar data point of cluster. The recall is the segment of the similar data points that are effectively clustered into the correct classes.

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

True positive is a number of relevant data instances in the dataset and false negative is number of non relevant instances in the dataset. Figure 3 illustrates the performance of the current mechanism on recall metric comparing it with conventional approaches.

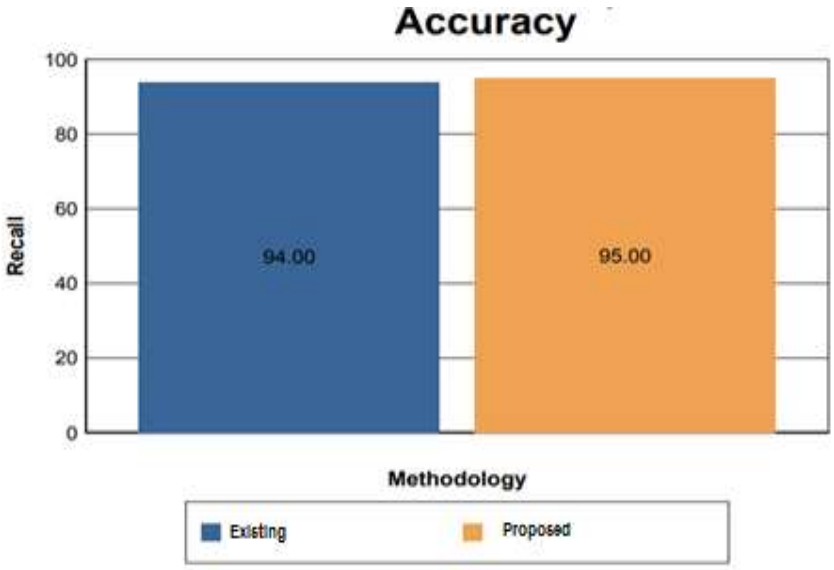

Figure 3: Performance analysis of the Recall Value

Cluster accuracy is based on activation function of the current approach. Convolution layer computes the feature map to produce subspace. F measure is a best metric to calculate the accuracy of the clustering mechanism. Although multiple data points may have different complications on cluster representation.

However, on certain data point, feature vector is minimized due to curse of dimensionality complexity. Otherwise, twitter data set yields the clusters which are reduced variance. current activation function acts as an clustering function to map the feature vector into a class distributions. Table 2 illustrates the performance of the proposed approach on cluster determination.

Table 2: Performance evaluation of Deep learning model with respect to Conventional mechanism to twitter Corpus dataset

S. No	Technique	Precision	Recall
1	Multiview Convolution Neural- Proposed model	97.80	95
2	K Means Clustering -Existing	97	94

In other words, the clustering architectures are highly efficient in computing clusters for streaming data and also efficient in computing the data representation of the data points distribution. Hyper-parameter is a vital component of the current deep learning models. In addition, the cross-validation is employed to twitter dataset to identify the good value of the hyper-parameters

Conclusion

Designed and implemented multiview convolution neural network Learning for unstructured and streaming data as deep learning architecture towards generating cluster

friendly representation. Proposed model uses the Convolution Neural Network with hyper parameter and loss function for high discriminating cluster after data pre-processing. Further deep learning mechanism gathers the optimal features in max pooling layer to illustrates feature for cluster generating using convolution, normalization and activation layer. Finally, SoftMax layer and loss layer is combined to yield the large variance clusters. Cluster performance evaluated utilizing f measure computation that it is capable in identifying the cluster. Finally current approach evaluates that it is high accurate and scalable on processing streaming text 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. Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 8, pp. 1798–1828, 2013
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. S.Praveen & R.Priya " A Deep Conceptual Incremental learning Based High Dimensional Data Clustering model- A Deep Learning Approach" in Turkish Journal of Computer and Mathematics, 2021
7. H. Liu, M. Shao, S. Li, and Y. Fu, "Infinite ensemble for image clustering," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, , pp. 1745–1754, 2016
8. Yookesh, T. L., et al. "Efficiency of iterative filtering method for solving Volterra fuzzy integral equations with a delay and material investigation." *Materials today: Proceedings* 47 (2021): 6101-6104.
9. 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
10. K. G. Dizaji, A. Herandi, C. Deng, W. Cai, and H. Huang, "Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization," in 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, , pp. 5747–5756, 2017.
11. Reddy, Ch Subba, T. L. Yookesh, and E. Boopathi Kumar. "A Study On Convergence Analysis Of Runge-Kutta Fehlberg Method To Solve Fuzzy Delay Differential Equations." *Journal of Algebraic Statistics* 13.2 (2022): 2832-2838.
12. 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.

13. 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.
14. 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.
15. W. Harchaoui, P. A. Mattei, and C. Bouveyron, ``Deep adversarial Gaussian mixture auto-encoder for clustering," in Proc. ICLR, 2017, pp. 1-5.
16. Navatha, S., et al. "Multitask Learning Architecture For Vehicle Over Speed As Traffic Violations Detection And Automated Safety Violation Fine Ticketing Using Convolution Neural Network And Yolo V4 Techniques." *Chinese Journal of Computational Mechanics* 5 (2023): 431-435.
17. N. Dilokthanakul et al. (2016). ``Deep unsupervised clustering with Gaussian mixture variational autoencoders." [Online]. Available: https:// arxiv.org/abs/1611.02648
18. G. Chen. (2015). ``Deep learning with nonparametric clustering." [Online]. Available: https://arxiv.org/abs/1501.03084