Journal of
Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.98549808

DEEP ADVERSARIAL REGULARIZED AUTOENCODER TECHNIQUE FOR HIGH DIMENSIONAL DATA CLUSTERING AND DATA PRIVACY PRESERVING PARADIGM IN BIG DATA CLOUD ENVIRONMENT

Kiruthika B[1], Dr. B. Srinivasan[2], Dr. P. Prabhusundhar[3]
[1 & 3]Assistant Professor in Computer Science
[2]Associate Professor in Computer Science
Gobi Arts & Science College (Autonomous), Gobichettipalayam, India.
[1]kiruthikabalu@gmail.com, [2]srinivasanb@gascgobi.ac.in
[3]drprabhusundhar@gascgobi.ac.in

Abstract

High dimensional data clustering have gained significant attention recently due to increased utilization of the high dimensional data streams across the large distributed cloud. High dimensional data clustering approaches has been developed in the conventional work using both machine learning and deep learning architectures. Despite of many advantageous, it is inefficient in handling the curse of dimensionality and data sparsity issues. Further learning model leads to high computational complexity and data disclosure attacks to various transaction queries to cloud servers. In order to mitigate those challenges, high scalable secure regularized model has been designed which is entitled as Deep Adversarial Regularized Autoencoder Technique. Autoencoder is a popular mechanism to accomplish dimensionality reduction. Proposed model deeply explore the latent structure of the data and computes the associations of the data points to construct the spatial and temporal cluster structures to high dimensional data as new clustering perspective. It discriminate the data points efficiently. Further evolving data streams are approximated using the variational autoencoder to preserve the cluster structures. Euclidean distance is employed in embedding function of the autoencoder to generate the efficient data clusters with minimized intra cluster similarity and inter cluster variation in the feature space. Hyper parameter tuning using RMSProp has been enabled in the output layer to make the data instance in the cluster to be close to each other by determining the affinity of the data on new representation. Experimental analysis has been performed on benchmarks datasets such as Twitter and Forest Cover to compute the proposed model performance with conventional approaches. The performance outcome represents that good scalability and effectiveness on high dimensional data has been reached.

Keywords: High Dimensional Data clustering, Big Data Cloud, Privacy Preserving, Variational Autoencoder, Advanced Networking

1. Introduction

Big Data Cloud environment is posing significant challenges in handling the distributed high dimensional data as it diverges greatly. Clustering is considered as an efficient data discriminative solution for rapid exploration of the data streams in the large distributed cloud[1]. Recently, high dimension data clustering has attained large research attention due to its increased dimension of the data. Traditional machine learning model has been employed to cluster the high dimensional data for valid representation using density based clustering[2], subspace clustering[3] and graph based clustering[4]. It captures the structural and contextual information of the data streams but it is inefficient in/ generalization and convergence of the data. In addition, those approaches are not sufficient in mitigating the data disclosure attacks to various transaction queries to cloud servers[5].

In order to manage those challenges, deep learning model has to be represented as high scalable secure regularized model. The proposed architecture entitled as Deep Adversarial Regularized Variational Autoencoder which is capable of handling linear and non linear feature learning to produce the high cluster friendly representation. Initially, missing value imputation and data normalization has been employed to pre-process the data. Pre-processed data is subjected to autoencoder model as it as popular mechanism to accomplish dimensionality reduction and extract the hidden structural information on the graph regularized data matrix. Proposed model deep explore the latent structure of the data and computes the associations of the data points to construct the spatial and temporal cluster structures to high dimensional data as new clustering perspective. It discriminate the data points efficiently.

Data disclosure attack[6] to the transactional queries towards data retrieval is secured on employing the strong encoding technique. Further evolving data streams are approximated using the variational autoencoder to preserve the cluster structures. Euclidean distance is employed in embedding function of the autoencoder to generate the efficient data clusters with minimized intra cluster similarity and inter cluster variation in the feature space. Hyper parameter tuning using RMSProp[7] has been enabled in the output layer to make the data instance in the cluster to be close to each other by determining the affinity of the data on new representation representing the hierarchical structure.

The article is sectioned into following section, literature similar to the high dimensional data clustering using machine learning and deep learning model has been analysed and described in detail is illustrated in section 2. The proposed architecture of deep adversarial regularized variational autoencoder is designed and implemented to secure data clustering in section 3 and performance evaluation of experimental outcomes using benchmark dataset produces high effectiveness of clustering model against conventional approaches is demonstrated in section 4. Finally article has been summarized in section 5.

2. Related work

In this part, high dimensional data clustering models in the big data cloud environment using deep learning and machine learning methods has been analysed in detail on its generated discriminate clusters along the processing of the data along the preprocessing and feature extraction process to generate the efficient clusters. Technique which performs similar to proposed model for high dimensional data clustering is described as follows

2.1. Anonymization based Deep Privacy Preserving Convolutional Autoencoder learning Technique

In this literature**,** deep privacy preserving convolutional auto encoder learning architecture for secure high dimensional data clustering on inferring the data distribution over the period. The proposed model leverages anonymization approach to secure the confidential and sensitive information. Anonymized data is subjected to feature learning along autoencoder for high cluster friendly representation on generation of objective function to produce maximum margin cluster. Those clusters are further fine tuned to feature refinement on the hyper parameter of various layers of deep learning model network to establish the minimum reconstruction error by feature refinement[8]. Softmax layer minimizes the intra cluster similarity and inter cluster variation in the feature space for cluster assignment. Hyper parameter tuning using stochastic gradient descent has been enabled in the output layer to make the data instance in the cluster to be close to each other by determining the affinity of the data on new representation.

2.2. Robust Subspace Clustering with Low-Rank Structure Constraint

In this literature, low rank structural model is designed to cluster the high dimensional data using robust subspace clustering. Initially data is represented in the affinity matrix which is further clustered using the rank the constraint. Extracted feature has been updated against the reconstruction error to obtain the salient features for clustering[9]. This model uses multiple layer to sparse representation of the salient features and used to eliminate the over fitting issue in the clusters by fine tuning of parameter with respect to certain criterion. On employing fuzzy rules, good representation of feature into cluster has been achieved on retaining the close affinity between the data instance in the cluster.

3. Proposed model

In this section, architecture of proposed deep adversarial regularized variational autoencoder model for high dimensional data clustering on inclusion of fine tuning of RMSProp of the encoder and decoder layers to generate the soft partition clusters on the big data cloud environment along privacy preserving has been illustrated as follows.

3.1.Data Pre-processing

Big Data cloud architecture handles the high dimensional data which contains the irrelevant attribute and noise attributes with missing value. Missing value prediction and data normalization is carried out using factor analysis[10] as it generates the normalized data. Data normalization is employed with Z score normalization[11].

- **Factor Analysis for Missing Value Imputation**

Factor analysis is employed for missing value imputation. Factor Analysis determines maximum common variance on the particular data field. It follows the Kaiser criterion which utilizes the Eigen value. It estimates the score for the variance of the particular data field to fill the missed value of the data field. It determines maximum likelihood value on basis of data correlation of the missing data field.

- **Variational Autoencoder - Dimensionality Reduction**

Variational Autoencoder[12] is applied to the preprocessed data as it capable of reducing the high dimensional data to low dimensional data on learning the dependencies and non dependency attributes. Variational autoencoder are feed forward acyclic neural network eliminate the non dependent attributes. Dependent attributes is represented as latent space in vector form containing the probability distribution.

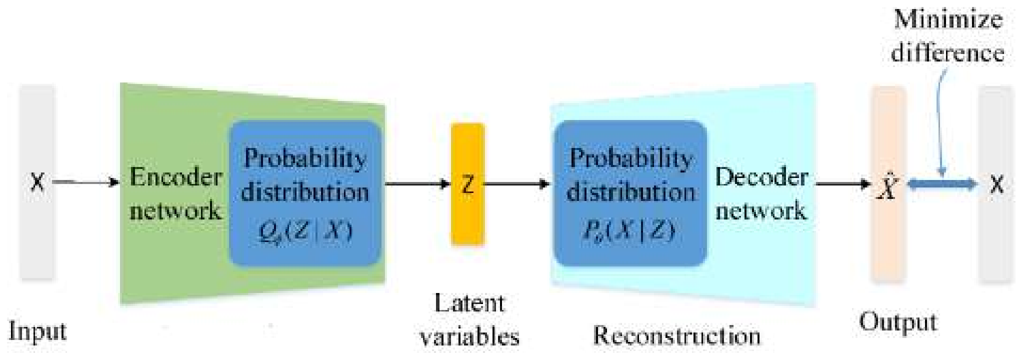

Figure 1: Autoencoder Based Dimensionality Reduction

Variational Autoencoder model considered as best transformation vector containing the encoder block, decoder block and latent spaces contain the latent variables. The optimal feature vector for clustering is extracted on reduction of the high dimensional feature space on employment of encoding process as latent space**.**

Original Input Data X= {Attribute$_1$, Attribute$_2$…….. Attribute}

Latent Space of the data L= {Attribute probability$_1$, Attribute probability$_2$… Attribute probability}

Encode Operation $Z = h(W_{encoding} X + b_{encoding})$

Where $W_{encoding}$ and $b_{encoding}$ is the parameter and h is the activation function

Optimal Reconstructed features = Decode(z)

Decode(z) = $h(W_{decoding}Z + b_{decoding})$

Where $W_{decoding}$ and $b_{decoding}$ is the parameter and h is the activation function

Probability distribution of the hidden attributes after decoding operation is given by KL divergence by

Probability distribution of attribute vector $P(z|x) = \frac{p(X|Z)p(z)}{p(x)}$

The variational attributes selected on inference to approximate the conditional probability distribution of the latent variables. KL divergence determines the disparity between the attributes and it minimizes the difference as best as possible. Model parameter and activation function on the probability distributions on the latent vectors with respect to loss function and shuffling of the feature reconstructed on decoding operation.

3.2.Deep Adversarial Regularized Autoencoder Clustering

Deep adversarial regularized variational autoencoder clustering is deep learning architecture and it is applied for clustering the compressed attributes of the input. In this part, variational autoencoder employs the fine tuning of the activation function with exponential linear unit towards cluster generation with objective functions of generative adversarial network[13].

- **Embedding loss function**

It includes the non linear loss components to eliminate the spatial and temporal structures of the clusters on embeddings. Embedding loss functions are latent representations of any two points which contains the maximum Euclidean distance and reduce the pair wise distance among the attributes values.

Embedding loss of the cluster structures = Min L(X, $X^{|}$) + $\sum_{k=0}^{n} \mathrm{G}(\mathrm{z})$

- **Activation layer**

Activation layer uses the ReLu function and exponential linear unit to produce the spatial and temporal cluster structures. It preserves the Inter-point distance. Activation layer of the model is capable of approximation of the inherent features of the original data. It is capable of computing the local properties of the data points. Regularization of the similarity matrix for hierarchical structure generation of cluster

Cluster structure of the data points is Min g(z) =$\frac{1}{2}\sum_{i<j}^{n} ||z_i||$

Table 1: Hyperparameter of the Variational Autoencoder

Hyper Parameter	**Values**
Cluster Batch Size	629
Autoencoder Learning Rate	0.04
Hidden layer	2 layers
Encoding Dimensions	50
Dataset Dimension	45
Epoch	50
Dropout	0.2
Optimizer	RMSprop
Loss Function	Poisson Function

- **Drop Out Layer**

The dropout layers eliminate the temporal feature which contains the transaction queries to the cloud database. Drop layer set the value of 0.2 which eliminate the overfitting issues and data disclosure attack. The adversarial regularized autoencoder model regularizes the temporal and irregular cluster structures. It is highly effective in discriminating the feature of the clusters. It contains the privacy preserving constraints to high level features.

High level feature is given as W_{ij} = { exp($\frac{x_i - x_j}{\alpha^2}$}

Privacy preserving constraints to high level feature is given as

$$P_{ij} = \begin{cases} 1 \\ 0 \end{cases} \quad \text{1 represent high level and 0 represent low level}$$

The privacy preserving approach heuristically computes the weight of the feature and represents it as low level to large abstract features and learns the discriminative features with few parameters. Pre-training the parameters has been used varied settings, while entire model is trained in an in fixed settings procedures for enabling faster training process. .

- **Optimizer function**

In this work, RMSProp is used as optimizer to increase the feature discriminant and overfitting issue of the cluster containing data points. It is considered loss reduction function. It contains high learning rate and it has good convergence rate. It reaches the global minima on the cost function with least possible value. High cluster representation depends on the weight and bias value of the gradient for the cluster. Gradient is given as

Cluster Gradient $G_w = \beta\, w + (1-\beta)w$

Cluster gradient minimize the distance of the data points in the cluster containing high level features of the data on effective decomposition of the feature or transfer of the feature to the outlier cluster as approximation.

- **Epoch Layer**

It is considered as number of iterations for the high level cluster structure. It determines the effective convergence of the model. Epoch layer generate the final weight of the similarity computation. It correlates the features into clusters. It learns the deep structure correlation on the adjustable parameters of the batch sizes.

Algorithm 1: Deep Adversarial regularized Variational Autoencoder

```
Input: High Dimensional Dataset
Output: High Representative Data Clusters
Process
  Data Pre-process ()
    Compute Attribute containing Missing value ()
         Factor analysis ()
              Kaiser Criteria on the attributes to insert the value of the missing field
    Dimensionality Reduction ()
       Feature Extraction_Autoencoder ()
   Encoder ()
          Assign Latent Variables
           G= {probability of the features)
                Encoded feature Set $F_s$ containing high level features
       Decoder ()
              Z= {encoded(g)}
                Decoded feature Set $F_s$ containing high level features
Apply Deep Learning of GAN ()
      Transfer learning ()
         Activation function ()
            Local properties of the data points
                Embedding Loss Layer ()
      Eliminate spatial and temporal structures of the cluster
         Dropout Layer ()
            Privacy preserving of the high level feature on eliminating the overfitting features
               Epoch()
                  Output Layer ()
                      Softmax ()--- Representation of the Cluster
              High Representative Cluster
```

Algorithm produces the cluster with the feature points as high representative cluster. It establishes the cluster with more weighted and minimized reconstruction error and clustering losses on the various dimensions of the data.

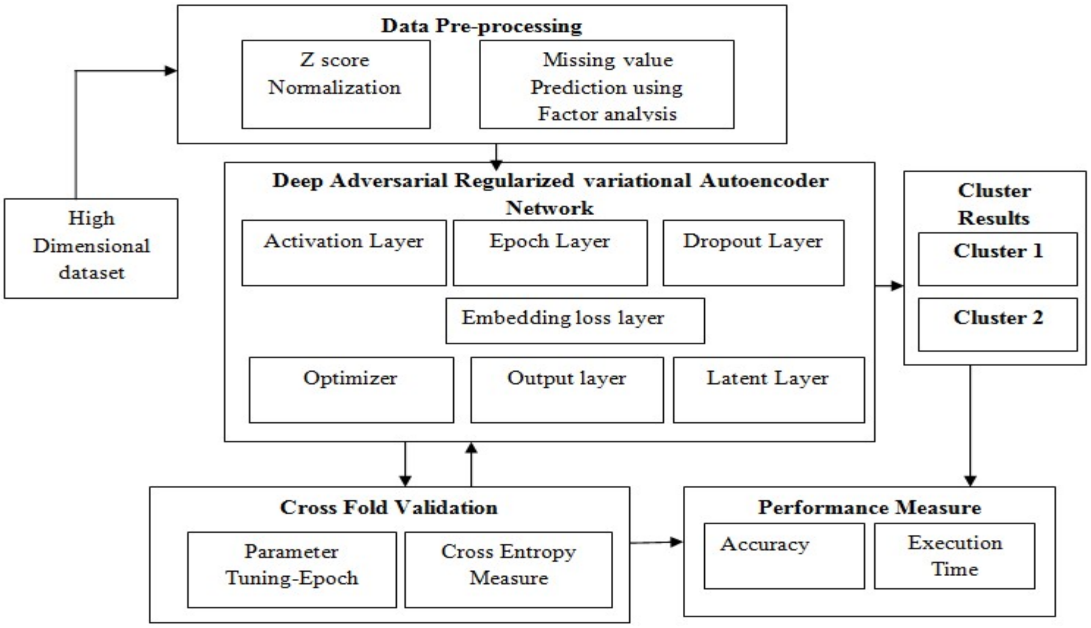

Figure 2: Architecture of the deep adversarial regularized clustering learning approach

Figure 2 represents the deep adversarial regularized variational autoencoder model for clustering of the high dimensional data. Model incorporates privacy preserving on encoding operation to secure the high level features. Activation layer uses the exponential linear unit to compute the high representative clusters. Proposed architecture guarantees produces high convergence and learning rate of the cluster generation. However, data points of the feature space are fine tuned using stochastic gradient decent to yield the better results.

Model is capable of managing the non linear dependency of the data points and utilizes the transfer learning in the activation function. Finally it avoids curse of dimensionality and the data sparsity challenges effectively on updating of the feature weight. Variational autoencoder learning algorithm has been employed to produce the cluster with high interclustre distance and maximizes the accuracy of clustering in addition to minimizing the reconstruction error on optimizing loss function.

On increase of the training epoch, the reconstruction error and clustering loss will slowly reduced and the accuracy of approach is enhanced [14]. Hence, weight updates details of the training phase of the variational autoencoder in the GAN have been detailed and Cross validation has been carried out on the test data to validate the model.

4. Experimental Results

Experimental analysis of variational autoencoder architecture for high representative cluster formation has been formulated out on the twitter dataset especially in CSV pattern for the data on the covid -19 pandemic[15]. Proposed performance has been evaluated on measure such as precision, recall and Fmeasure. In this work, 60% of twitter dataset is subjected to training and 20% of dataset is subjected to clustering and remaining 20% of dataset is to validate the proposed model on cross fold validation.

Finally performance of the proposed model is cross evaluated employing the 10 fold validation. Figure 2 represents the performance evaluation of the proposed architecture in terms of precision on twitter dataset. The training parameter of the comprehensive encoder learning has been specified in the table 2

Table 2: Training parameters

Parameter	**Value**
Cluster Learning rate of the data	10^{-6}
Loss Function	Poisson Function
Batch size	15
Epoch	45

.

4.1.Dataset description : Twitter

Twitter data set contains 340,000 Twitter messages (tweets) of various health trajectories on wide classes in various area and subjects on world in the covid -19 pandemic period. Data is represented in CSV format.

4.2. Performance Evaluation

The proposed architecture has been computed using confusion matrix on 10 fold validation. In this work, proposed model compute the performance of cluster intercluster distance similarity to on dataset mentioned. The performance evaluation of the deep learning model depends on the process of activation function, dropout layer, embedding loss layer, optimizer function and hyper parameter of model. In order to privacy preserve the transactional queries, rank constraints for high level features considered as criteria to eliminate the overfitting issue and outlier cluster formation in the output layer of the model.

- **Precision**

It is a measure of positive predictive value. It is further illustrated as the fraction of relevant data points among the each cluster structures projected employing the model. In other words Precision is termed as ratio of number of optimal feature to the number of all subspace of the feature set extracted. Figure 2 represents the performance evaluation of the proposed architecture in terms of precision on twitter dataset. 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.

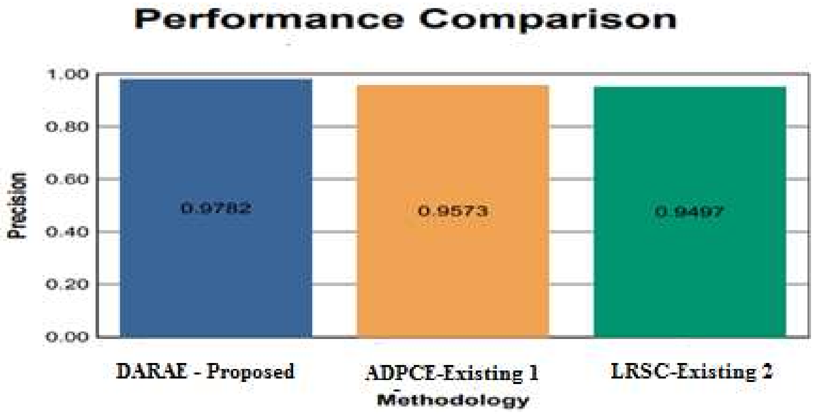

Figure 2: Performance Evaluation of the methodology with respect to Precision

.

- **Recall**

Recall is the part of similar data points in the cluster which is extracted from the decoder part of the model to the total amount of similar data point gather for entire cluster. The recall is the considered of the relevant feature that are clustered.

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

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

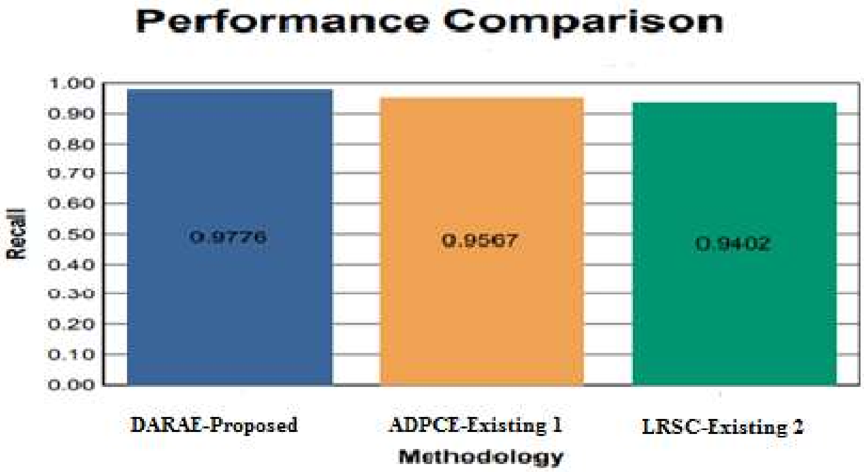

Figure 3: Performance Evaluation of the methodology with respect to Recall

Quality of the data cluster depends on activation function and embedded loss layer of the model. Encoder dimension calculates the high level features to create subspace and dropout layer minimize the low level feature of the data. F measure is a measure of the cluster quality using epoch and batch size of the cluster on eliminating the overfitting issue and efficient in generating the outlier to the data points.

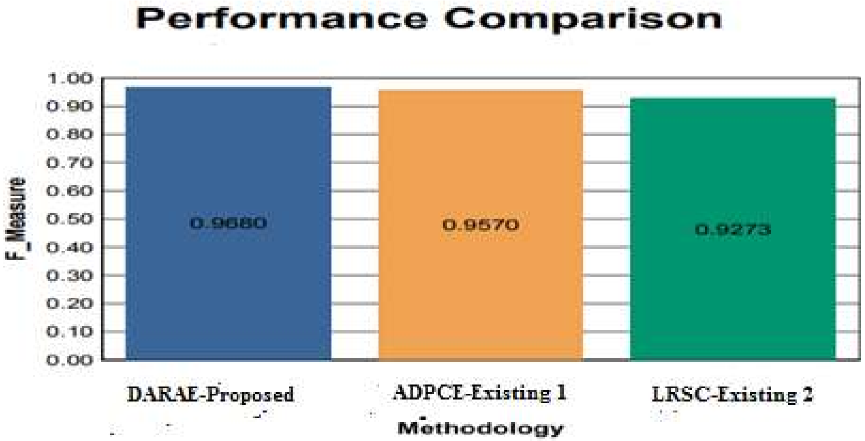

Figure 4: Performance Evaluation of the methodology with respect to F- Measure

- **F measure**

It is the number of relevant data points in the clusters to streaming data in the learning model. It is considered as accuracy. Accuracy is given by

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

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 autoencoder architecture against conventional approaches

Technique	Precision	Recall	F measure	Accuracy
Deep Adversarial regularized variational autoencoder model - Proposed	0.9782	0.9776	0.9680	0.9921
Anonymization deep preserving convolution autoencoder –Existing 1	0.9573	0.9567	0.9570	0.9847
Low rank structure constraint for subspace clustering- Existing 2	0.9497	0.9402	0.9273	0.9448

Finally, proposed clustering approach not only identifies high representative clusters to the given high dimensional data, but it is capable of securing the high level feature of the data against the data disclosure attacks to big data cloud environment.

Conclusion

Deep Adversarial Regularized Autoencoder technique is designed and implemented in this work to generate high representative clustering. On processing the proposed model, it is capable of exploring the deep explore the latent structure of the data and computes the associations of the data points to construct the spatial and temporal cluster structures to high dimensional data as clustering perspective. Further evolving data streams are approximated using the variational autoencoder to preserve the cluster structures. Euclidean distance is employed in embedding function of the autoencoder to generate the efficient data clusters with minimized intra cluster similarity and inter cluster variation in the feature space using the RMSProp based hyper parameter tuning in the output layer to make the data instance in the cluster to be close to each other. Finally proposed model proves that it is effective and high scalable on high dimensional data on achieving the high accuracy in the generated cluster structures.

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