Outlier Detection in High Dimensional Data

Anusha L, Nagaraja G S

***Abstract:** Artificial intelligence (AI) is the science that allows computers to replicate human intelligence in areas such as decision-making, text processing, visual perception. Artificial Intelligence is the broader field that contains several subfields such as machine learning, robotics, and computer vision. Machine Learning is a branch of Artificial Intelligence that allows a machine to learn and improve at a task over time. Deep Learning is a subset of machine learning that makes use of deep artificial neural networks for training. The paper proposed on outlier detection for multivariate high dimensional data for Autoencoder unsupervised model.*

***Keywords:** Outlier Detection, Autoencoder Model, Unsupervised Model.*

I. INTRODUCTION

Outlier Detection is a data point deviates from the normal points. In statistics-based intuition it has normal and abnormal mechanism. The scenario in outlier detections is supervised, unsupervised, semi-supervised model. Supervised model in which data has known labels or output, used in insurance underwriting, fraud detection. Unsupervised model in which labels or output unknown focus on finding patterns and gaining insight from the data, used in customer clustering, associate rule mining. Semi-Supervised model in which labels or output known for a subset of data. A combination of supervised and unsupervised learning is utilized in medical forecasts.

Application of outlier detection are intrusions in communication networks, fraud in financial data, fake news and misinformation, health care analysis, industry damage detection, security, and surveillance. The types of errors found in outlier are data entry errors its type of human errors, measurement errors it is an instrumental error, experimental error, intentional error, data processing error, sampling errors, natural errors. The impact of outlier may result in change in the results, error variance, normality, bias, assumption.

Methods of outlier detection are extreme value analysis is the one dimensional data value which are too large or the small are the outlier detection, probabilistic and statistical modelling which assume specific distribution for the data in which low probability of the members are marked as the outlier, proximity based models contains cluster based method in which it classify data in to different clusters and point that as out of cluster in cluster based method, High dimensional data which is used for multivariate variable it contains gaussian distribution, autoencoder model.

II. SYSTEM ARCHITECTURE

In figure shows the system architecture of autoencoder model, an autoencoder neural network is an unsupervised machine learning algorithm that applies back propagation, setting the target values to be equal to the inputs [1]. Key facts about autoencoder it is an unsupervised machine learning algorithm similar to principal component analysis, it minimizes the same objective function as PCA, it is a neural network in which the neural networks target output as its input.

Types of autoencoders are convolution autoencoder which contains image deconstruction, image colorization, advanced application[2]. The spare autoencoder is the next type, which provides an alternate approach for introducing a bottleneck without reducing the number of nodes in hidden layers.

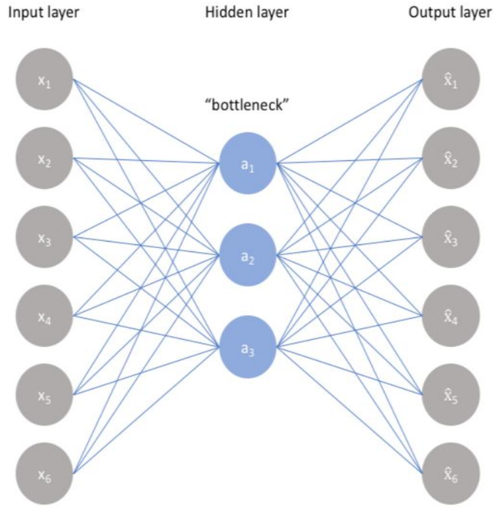


Figure 1: Autoencoder model

A) Encoder: This is the network components that compresses the data into latent space representation.
B) Code: This is the part of the network represents the compressed input that is fed into decoder.
C) Bottleneck: It is an approach for deciding which aspects of observed data are relevant information, compactness of representation measured as the compressibility.
D) Decoder: The goal of this section is to reassemble the input from the latent space representation.

Manuscript received on May 22, 2021.
Revised Manuscript received on May 26, 2021.
Manuscript published on June 30, 2021.
* Corresponding Author

Anusha L*, Student, Department of Computer Science and Engineering, Rashtreeya Vidyalaya college of Engineering, Bengaluru, (Karnataka), Email: anushal.scn19@rvce.edu.in

Dr. Nagaraja G S, Professor and Associate Dean, Department of Computer Science and Engineering, Rashtreeya Vidyalaya College of Engineering, Bengaluru (Karnataka), India. Email: nagarajags@rvce.edu.in

III. METHODOLOGY

Autoencoder model for anomaly detection contains:

A) Import all required libraries for autoencoder model: For anomaly detection in autoencoder model contains tensorflow, keras as backend, seaborn as data visualization.

B) Load the datasets: The datasets have been imported using csv file so that interactive chart has been ploted. The datasets used in the models can be for anomaly detection, bank fraud detection, intrusion detection system, medical analysis.

C) Data Preprocessing: Involves standardize the target vector by removing mean and scaling into unit variance we use standard scalar function.

The total datasets have been divided in to training data frame and testing data frame and printed the values then apply standardization for the datasets.

D) Create training and test splits splits the data into subsequence, reshape the input data into a shape, no of sample, no of timesteps, no of feature.

E) Build an LSTM Autoencoder model: Initially train the autoencoder with no anomalies then take new data point and reconstruct it using autoencoder, if reconstruct error point above some thresholds point, we set that datapoint as an anomaly, it contains encoder, decoder and repeat vector. Repeat vector layer duplicated the output from our encoded representation.

F) Train the autoencoder: the autoencoder model has been trained to know the validation loss, training loss for the datasets.

G) Plot metrics and evaluate the model: In graph validation loss will be consequently lower than training loss meaning underfit the data, high dropout values are used. To detect anomaly, calculate mean absolute error for the training data. Then set the threshold value for the datasets if the threshold value is greater than the value is anomaly.

H) Detect anomalies for the datasets: Based on the threshold value the anomaly has been detected. The threshold value is less then test loss there is no anomaly if it greater then test loss there will be anomaly in the network.

IV. RESULTS AND ANALYSIS

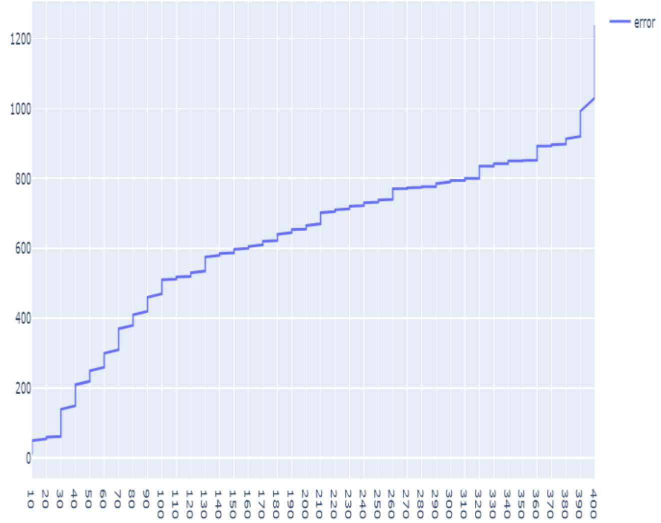
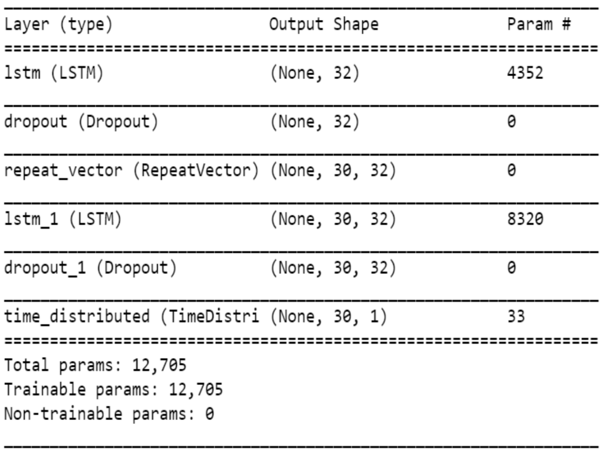


Figure 4.1: Interactive Line Chart



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm (LSTM)                  (None, 32)                4352
_________________________________________________________________
dropout (Dropout)            (None, 32)                0
_________________________________________________________________
repeat_vector (RepeatVector) (None, 30, 32)            0
_________________________________________________________________
lstm_1 (LSTM)                (None, 30, 32)            8320
_________________________________________________________________
dropout_1 (Dropout)          (None, 30, 32)            0
_________________________________________________________________
time_distributed (TimeDistri (None, 30, 1)             33
=================================================================
Total params: 12,705
Trainable params: 12,705
Non-trainable params: 0
_________________________________________________________________
```

Figure 4.2: Build Autoencoder Model

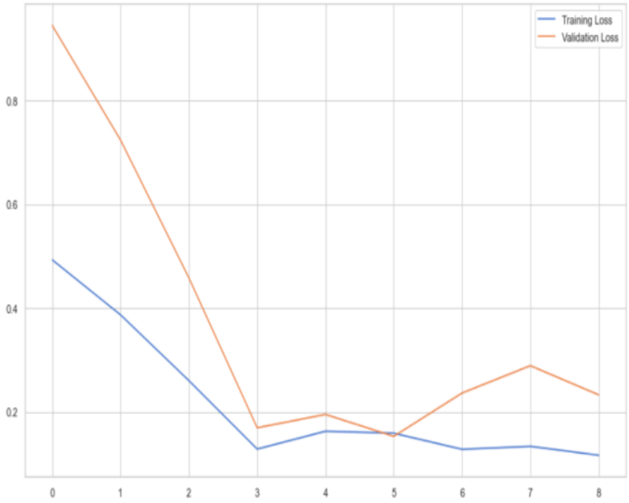


Figure 4.3: Plot the graph and evaluate the model.

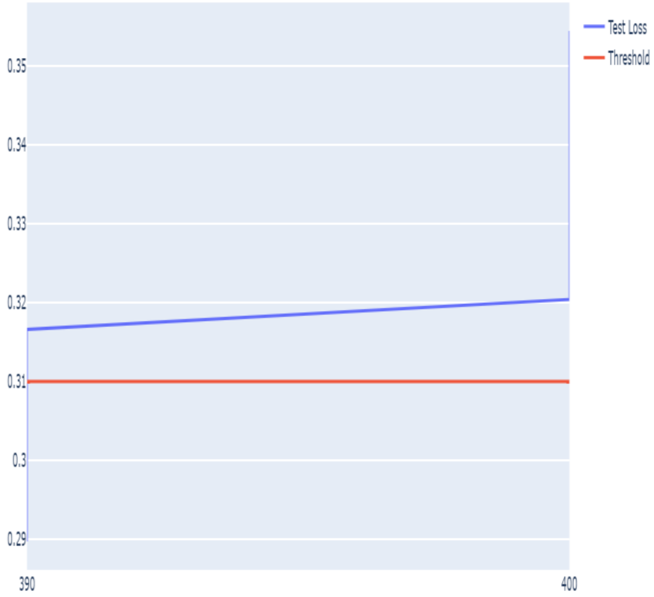


Figure 4.4: Threshold value graph

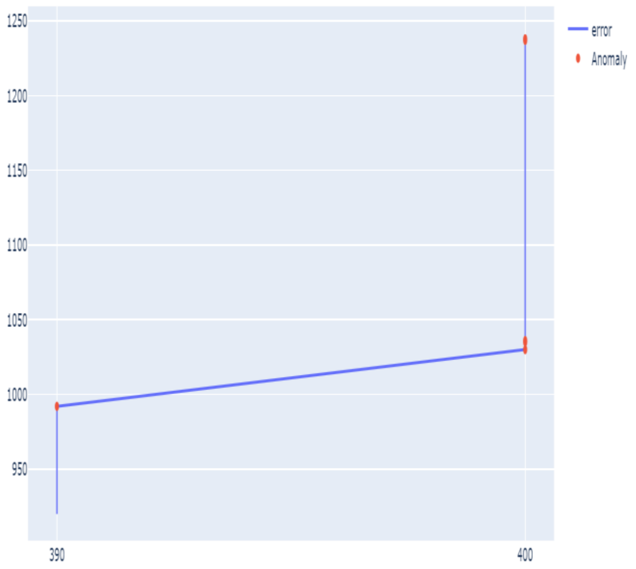


Figure 4.5: Anomaly Detection

In figure 4.1 shows the interactive line charts for the data set model for multivariate high dimensional data in the format of time series. In figure 4.2 shows the autoencoder model has been build which has encoder, decoder and repeat vector. Training of autoencoder model will happens in this stage. In figure 4.3 shows the graph of validation loss and training loss based on the dataset that has been imported. In figure 4.4 shows the calculation of threshold value based on validation loss of the datasets. In figure 4.5 shows the anomaly or outlier detection for the autoencoder model, there is a variation from the normal values in the datasets.

V. CONCLUSION

The paper provides implementation details of Autoencoder unsupervised artificial neural network for high dimensional data, in which the neural networks target output as its input. Autoencoder are popular choices for outlier detection, so Autoencoder model is used in image compression, denoising, generation, dimensional reduction, feature extraction, sequence prediction.

REFERENCES

1. K. Han, Y. Wang, C. Zhang, C. Li and C. Xu, "Autoencoder Inspired Unsupervised Feature Selection," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 2941-2945, doi: 10.1109/ICASSP.2018.8462261.
2. N. Hallett et al., "Deep Learning Based Unsupervised and Semi-supervised Classification for Keratoconus," 2020 International Joint Conference on Neural Networks (IJCNN), 2020, pp. 1-7, doi: 10.1109/IJCNN48605.2020.9206694.
3. M. S. Kim, J. P. Yun, S. Lee and P. Park, "Unsupervised Anomaly detection of LM Guide Using Variational Autoencoder," 2019 11th International Symposium on Advanced Topics in Electrical Engineering (ATEE), 2019, pp. 1-5, doi: 10.1109/ATEE.2019.8724998.
4. O. I. Provotar, Y. M. Linder and M. M. Veres, "Unsupervised Anomaly Detection in Time Series Using LSTM-Based Autoencoders," 2019 IEEE International Conference on Advanced Trends in Information Theory (ATIT), 2019, pp. 513-517, doi: 10.1109/ATIT49449.2019.9030505.
5. Z. Li et al., "Unsupervised Clustering through Gaussian Mixture Variational AutoEncoder with Non-Reparameterized Variational Inference and Std Annealing," 2020 International Joint Conference on Neural Networks (IJCNN), 2020, pp. 1-8, doi: 10.1109/IJCNN48605.2020.9207493.
6. T. Peken, R. Tandon and T. Bose, "Unsupervised mmWave Beamforming via Autoencoders," ICC 2020 - 2020 IEEE International Conference on Communications (ICC), 2020, pp. 1-6, doi: 10.1109/ICC40277.2020.9149222.

AUTHORS PROFILE

Anusha L, is a MTech student at Department of Computer Science and Engineering, Rashtreeya Vidyalaya college of Engineering, Bengaluru, Karnataka. anushal.scn19@rvce.edu.in

Dr. Nagaraja G S, is working as professor and Associate Dean at Department of Computer Science and Engineering, Rashtreeya Vidyalaya College of Engineering, Bengaluru, Karnataka, India. nagarajags@rvce.edu.in