PalArch's Journal of Archaeology of Egypt / Egyptology

A STUDY ON OUTLIER DETECTION USING HYBRID CONVOLUTION NEURAL NETWORK AND NOVEL BI-DIRECTIONAL GATED RECURRENT UNIT (CNN-BIGRU)

G.Manoharan[1], K. Sivakumar[2]*

[1*]Research Scholar,Department of Mathematics,Sathyabama Institute of Science andTechnology,Chennai,Tamilnadu,India.

[2]Professor,Department of Mathematics,Sathyabama Institute of Science and Technology,Chennai,Tamilnadu,India.

[1*]vijimanoharan77@gmail.com,[2]siva111k@gmail.com

G. Manoharan, K. Sivakumar.A Study on Outlier Detection Using Hybrid Convolution Neural Network and Novel Bi-Directional Gated Recurrent Unit (CNN-BIGRU)-- Palarch's Journal of Archaeology of Egypt/Egyptology 17(7), 4802-4808. ISSN 1567-214x

Keywords: Outlier Detection, Convolution Neural Network, Bi-directional Gated Recurrent Unit (Bi-GRU), F-score, AUC.

ABSTRACT:
In this paper a study onConvolution Neural Network and novel Bi-Gated Recurrent Unit (Bi-GRU) with extraction stacked auto encoders are discussed, further the data using outliers and inliers have been classified by Probabilistic Neural Networks (PNNs).This kind of system shows better performance compared with the existing technologies. More hidden layers/depth usage in deep learning framework resulted in better performance. However too much of depth may be avoided for better accuracy and hence this current work can be extended by CNN-BiGRU and evaluated in terms of the efficiency, F-score, AUC for the outlier detection.

INTRODUCTION

An outlier is considered as data point which is substantially varied from the rest of the data. The outlier is stated as, "An outlier is generates by various mechanism as it is considered as an observation differentiates from other observations which in turn raised into suspicious". Hence the outliers are otherwise termed as anomalies, discordant, deviants and abnormalities presented in the various data. In most of the applications the generated data by the creating process reflecting the system activity or the entities collections by the observations made. The outliers are created when the generating process can performs unusually. Hence the abnormal characteristics of the system and data generation process effect presented in the outliers which also termed as

useful information. the application-specific insights which is useful are generated from the unusual characteristics recognition[1, 2].

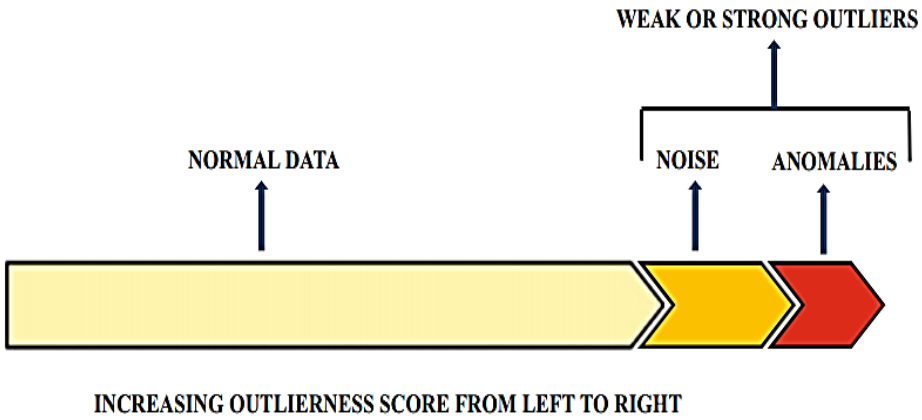

Figure1. Example -Weak/Strong Outliers[1]

Basically the outliers of data point measures by most of the detection techniques[3, 4] which used some quantified measures like nearest neighbour based distance and underlying region's sparsity. On a continuous spectrum each data point is presented as shown in Fig.1 from normal data proceeded to noise and anomalies termed as a weak or strong outliers.Both techniques are exhibiting better performances in their exiting work[5, 6]. In this study, the efficient techniques is used for detecting the outliers considered as the Convolution Neural Network (CNN) and Bi-directional Gated Recurrent Unit (Bi-GRU).

OBJECTIVE

The main objective of this study includes,

- To perform feature selection effectively based on property value.
- To identify the outliers by using the hybrid approach of Convolution Neural Network and new Bi-GRU (Bi-directional Gated Recurrent Unit).
- To evaluate the current study for the outliers detection accuracy.

METHODOLOGY

For effective normal data differentiation from the outliers and to detect the outliers accurately the new hybrid approach proposed as shown in proposed flow diagram Fig. 2.

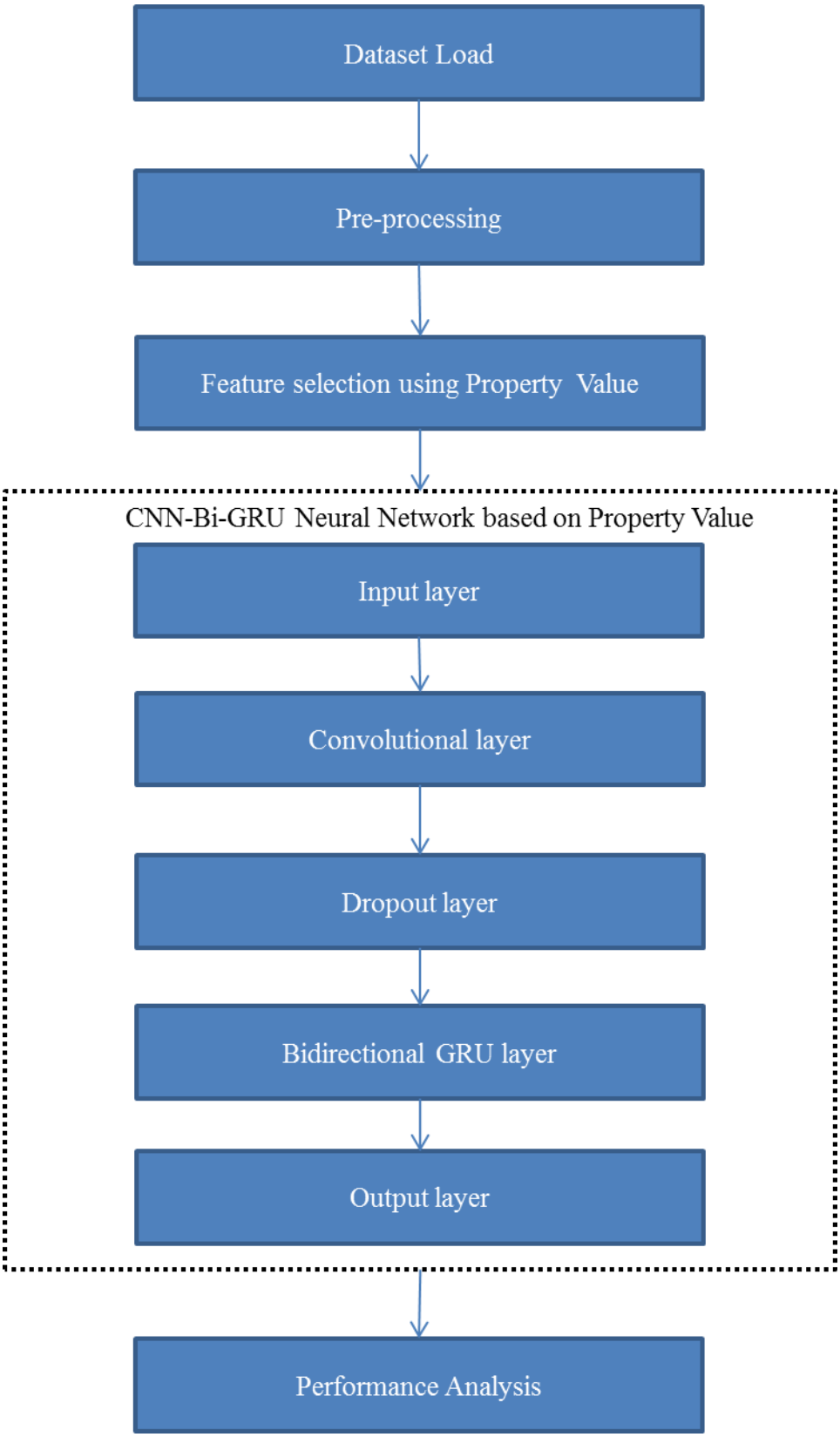

Figure 2. Proposed Flow

The proposed flow shows the overall process of our current study as,

- Initially, the type of data and nature of data are selected for the current study.
- The efficient feature selection are performed for the current selected data using the new property value technique which selects only the best value and leave the remaining values.
- Further the best feature selected data are given as input to the hybrid system.
- The stacking process used to increase the outlier detection and accuracy improved for the outlier data identification using the hybrid approach.
- The CNN algorithm and Bi-GRU followed for the effective classification

CNN and Bi-GRU Algorithm

- For better classification CNN is significantly designed. CNN can also be used as a time series analysis for 1-D data. For higher performance on the non-linear problems the CNN exhibits the weight sharing concept.
- The convolution and dropout layers of internal operation is shown in above figure. On input data the convolution is applied and transformed to feature maps. Followed by dropout layer which is used to model the obtained feature maps and best features are converted.
- For spatial feature extraction CNN is used from the input data and finally the processed data transformed into Bi-GRU unit.
- The best suited outlier detection is selected by property type.

STATEMENT OF THE PROBLEM

Based on machine learning techniques, this study reviewed about the outlier detection methodologies for IoT.This method illustrated about the two network frameworks used for outlier detection. Distributed and centralized measured approaches have approved in planar network framework. The centralized structure for outlier detection exhibited main processing time delay and greater energy consumed due to the end nodes data transmission as multi-hop to data center takes more time and also greater communication overhead resulted. In distribution system the outlier detection performed by each node and however resource constrained seen in end nodes with the restricted computing capacity, energy level and memory space. In the hierarchical network framework, the fog computing scheme performed well whereas cluster based approach reduce the computational overhead and put the burden on the end sensor nodes. This study also suggested that compared with single machine learning algorithm ensemble learning suitable to detect any kind of data's outliers yet it shows difficulty in adoption on end sensor nodes[7]. Another study using the combination of CNN and GRU a hybrid sequential learning based energy forecasting system has been developed. Energy consumption reduced. Due to the CNN feature extraction potential and GRU the proposed system reduced computational difficulty and accuracy is predicted. However, the lesser error rate has been resulted compared with existing studies[8].

In our study, the above research challenges are expected to overcome by the hybrid approach using the Convolution Neural Network and novel Gated

Recurrent Unit (Bi-GRU) to attain efficiency and greater accuracy in detecting the outliers.

MAJOR CHALLENGES IN OUTLIER DETECTION

- The locally abnormal behaviours are presented by the outliers. If it is the high dimensional data, the local correlation of features exploring and measurement is considered as difficult strategy.
- Likewise the real outliers the noise in the data are imitated and hence it is another challenge to differentiate it.
- Implementing the combining approach to choose the right subspace or learning bases for detecting the outliers.
- Differentiating the normal behaviours and outliers from the different data is the trial and evaluating the outliers also another issue.
- The other general complexities are termed as input data nature, type of outliers and data labels and consumption of memory and time.

SCIENTIFIC BACKGROUND

For credit scoring, a better predictive power is obtained through improved outlier adaptation with new multi-stage ensemble model has been proposed in this study. In the noise-filled datasets the outliers opposing impacts reduced by the algorithm using the bagging approach to recognize the outliers and in turn the outlier adapted training set has been constructed and base classifiers of outlier adaptability is improved further.In this study the predictive power improved by the ensemble learning based on stacking technique for parameter optimization which is self-adaptive[9].Different correlation of human behaviour information/outliers in smart cities has identified by knowledge discovery and data mining considered as first algorithm. Followed by the collection of abnormal human behaviour has been identified by the deep convolution neural networks in which various historical data features learned are considered as second algorithm. In terms of accuracy the deep learning framework has been performed well compared with data mining techniques[10]. For the unsupervised outlier detection in time series, this studysuggested based on the sparsely connected Recurrent Neural Network (RNN) two auto-encoder ensemble scheme have been proposed. Effective performance resulted from this study. Various auto encoder design and latest deep neural network structures can be used in future for accuracy enhancement are suggested in this work[11].

For the Wi-Fi indoor localization environment the if-Ensemble technique used for outlier detection with Ensemble learning, supervised and unsupervised learning combination. For initial classification of normal and abnormal data the unsupervised learning method of isolation Forest (iForest) has been used. For the enhanced separation of normal and outliers the stacking based Ensemble learning has been used[12]. The Ensemble based Partial Least Squares (EnPLS) framework has been presented for the outlying molecules detection and informative descriptors selection followed by the proposed model improvement. EnPLS ability in different prediction tasks have been established by the aqueous solubility data analysis. The data analysis also facilitated by the proposed EnPLS technique[13].For anomaly detection the

mixture of expert's model has been proposed to assess the larger databases. The outliers are identified by this proposed model and many if the individual algorithms have not considered. The empirical assessment as the number of dimensions in information raises resulted with linear scaling and better scalability[14]. The Gated Recurrent Unit (GRU) and 1D- Convolution Neutral Network (CNN) have been used in this study to learn the controllers and sensor parameter values dependencies and the spatial temporal correlation for the anomaly detection in industrial control systems. The training complexity occurs if the larger number of model parameters are seen. In learning the relationship among the networks used in data features optimization. The rule based detection has been followed[15].

RESEARCH GAP

For an effective outlier detection a sequential ensemble-based Sparse Modelling has been proposed. For high dimensional numerical data with noisy features the proposed technique performed outlier scoring, feature selection with iterative mutual refinement and finally reliable outlier scores has been obtained. The dependent base models have been constructed by sequential ensembles with the help of recent base model to improve the following one. The major disadvantage was complexity in generating sequential ensembles to detect the outliers as the class labels which are rumoured as unavailable[16]. To detect the outliers the cross-validation is also used to classify the true normal sample from the true outlier sample. In this study Ensemble Learning based Outlier detection (ELO) has been developed. However the danger exists that in some cases the normal samples have mistakenly considered as outlier samples and hence detection in outliers have not shown accuracy. To avoid this ensemble learning has been used in this existing study but it may be repeated to avoid this misconception[17].

CONCLUSION

In the current study, for the feature selection based on property value method is followed. The new hybrid approach is presented for the outlier detection which comprised of CNN and new Bi-GRU (Bi-directional Gated Recurrent Unit). The proposed method CNN-BiGRU is evaluated in terms of the efficiency, F-score, AUC for the outlier detection from various data. It is expected that the higher efficiency is resulted by compared with the existing approaches used for outlier detection.

REFERENCES

Aggarwal, C.C. (2017). An introduction to outlier analysis. In *Outlier analysis* (pp. 1-34). Springer, Cham.

Wang, H., Bah, M.J., & Hammad, M. (2019). Progress in outlier detection techniques: A survey. *IEEE Access*, *7*, 107964-108000.

Wahid, A., & Rao, A.C.S. (2020). ODRA: an outlier detection algorithm based on relevant attribute analysis method. *Cluster Computing*, 1-17.

Sainis, N., Srivastava, D., & Singh, R. (2018). Feature classification and outlier detection to increased accuracy in intrusion detection system. *International Journal of Applied Engineering Research*, *13*(10), 7249-7255.

Čeponis, D., & Goranin, N. (2020). Investigation of Dual-Flow Deep Learning Models LSTM-FCN and GRU-FCN Efficiency against Single-Flow CNN Models for the Host-Based Intrusion and Malware Detection Task on Univariate Times Series Data. *Applied Sciences*, *10*(7), 2373.

Fanta, H., Shao, Z., & Ma, L. (2020). SiTGRU: Single-Tunnelled Gated Recurrent Unit for Abnormality Detection. *Information Sciences*.

Jiang, J., Han, G., Shu, L., & Guizani, M. (2020). Outlier Detection Approaches Based on Machine Learning in the Internet-of-Things. *IEEE Wireless Communications*, *27*(3), 53-59.

Sajjad, M., Khan, Z. A., Ullah, A., Hussain, T., Ullah, W., Lee, M. Y., & Baik, S. W. (2020). A novel CNN-GRU-based hybrid approach for short-term residential load forecasting. *IEEE Access*, *8*, 143759-143768.

Zhang, W., Yang, D., Zhang, S., Ablanedo-Rosas, J. H., Wu, X., & Lou, Y. (2020). A novel multi-stage ensemble model with enhanced outlier adaptation for credit scoring. *Expert Systems with Applications*, *165*, 113872.

Belhadi, A., Djenouri, Y., Srivastava, G., Djenouri, D., Lin, J. C. W., & Fortino, G. Deep learning for pedestrian collective behavior analysis in smart cities: A model of group trajectory outlier detection. *Information Fusion*, *65*, 13-20.

Kieu, T., Yang, B., Guo, C., & Jensen, C. S. (2019). Outlier Detection for Time Series with Recurrent Autoencoder Ensembles. In *IJCAI* (pp. 2725-2732).

Bhatti, M.A., Riaz, R., Rizvi, S.S., Shokat, S., Riaz, F., & Kwon, S.J. (2020). Outlier detection in indoor localization and Internet of Things (IoT) using machine learning. *Journal of Communications and Networks*, *22*(3), 236-243.

Cao, D.S., Deng, Z.K., Zhu, M.F., Yao, Z.J., Dong, J., & Zhao, R.G. (2017). Ensemble partial least squares regression for descriptor selection, outlier detection, applicability domain assessment, and ensemble modeling in QSAR/QSPR modeling. *Journal of Chemometrics*, *31*(11), e2922.

Nun, I., Protopapas, P., Sim, B., & Chen, W. (2016). Ensemble learning method for outlier detection and its application to astronomical light curves. *The Astronomical Journal*, *152*(3), 71.

Xie, X., Wang, B., Wan, T., & Tang, W. (2020). Multivariate Abnormal Detection for Industrial Control Systems Using 1D CNN and GRU. *IEEE Access*, *8*, 88348-88359.

Pang, G., Cao, L., Chen, L., Lian, D., & Liu, H. (2018,). Sparse Modeling-Based Sequential Ensemble Learning for Effective Outlier Detection in High-Dimensional Numeric Data. In *AAAI* (pp. 3892-3899).

Kaneko, H. (2018). Automatic outlier sample detection based on regression analysis and repeated ensemble learning. *Chemometrics and Intelligent Laboratory Systems*, *177*, 74-82.