Artificial Intelligence and Social Computing, Vol. 122, 2024, 175–181
https://doi.org/10.54941/ahfe1004654

Application of Long Short-Term Memory (LSTM) Autoencoder With Density-Based Spatial Clustering of Applications With Noise (DBSCAN) on Anomaly Detection

Chauchen Torng and Hehe Peng

Dept. of Industrial Eng. & Mgmt, National Yunlin University of Science and Technology, Taiwan

ABSTRACT

This study explores the use of LSTM (Long Short-Term Memory) autoencoder combined with DBSCAN (Density-based spatial clustering of applications with noise) under the condition of data imbalance. The reconstruction error of the model after training is used as an evaluation index where the errors of each time point between the reconstruction sequence and the actual sequence are calculated and inputted for classification in the DBSCAN model. In this study, a water distribution system dataset from the SKAB (Skoltech Anomaly Benchmark) was used to verify the anomaly detection of our proposed model. Our model shows the F1-score of 0.8025 which is better than the four models proposed by Moon et al. in 2023. With a LSTM autoencoder, the proposed DBSCAN classification model can avoid the difficulty of setting a threshold value in classification.

Keywords: Anomaly detection, Condition-based maintenance, LSTM autoencoder, DBSCAN

INTRODUCTION

The manufacturing industry is facing digital transformation, and the number of equipment and production complexity of manufacturing systems are increasing, which is more likely than ever to cause failures and downtime in the manufacturing process, making equipment management more difficult. In order to avoid unplanned failures, equipment maintenance is required to avoid downtime in the production process. Anomaly detection is a binary classification problem, which needs to be carried out in the form of online or real-time detection, the collection of data points and the monitoring process run at the same time, and each data point can be processed at the same time when it is updated. Mao et al. (2022) indicates that online detection can help to monitor the change of equipment status in a short time and avoid economic losses caused by downtime, and online detection must have good real-time performance. The status of early equipment failures should be identified as quickly and accurately as possible, and the detection scheme should be robust enough to avoid false alarms.

LSTM autoencoder refers to the encoder and decoder in the autoencoder are composed of LSTM network, which can make this model structure support input sequence and output sequence, and can capture the time dependent multivariable data better than the traditional autoencoder, which is suitable for modeling time series data, and can be applied to sequence prediction, anomaly detection and sensor signal analysis. The use of LSTM autoencoders for multivariate time series data has been used in several studies. de Pater and Mitici (2023) used LSTM autoencoders to construct health metrics and RUL predictions for aircraft engines, and the proposed method reduces RMSE by 19%.

DBSCAN (Density-based spatial clustering of applications with noise) is a spatial clustering method based on density, which is one of the common clustering methods (Singh et al., 2022), which is an unsupervised learning method that does not require labels for data, classifying points with similar properties into the same group, without defining them, DBSCAN needs to give a neighborhood distance of ε and minPts (minimum points). The neighborhood distance is the radius distance of a point, and the point within the distance is the neighbor of that point; minPt is the minimum number of points required for a point to form a cluster with its neighbors, and if the number of neighbors at a point is more than minPts, this group of points can be called clustered. DBSCAN is applied to anomaly detection, in Çelik et al. (2011); Garg et al. (2020); Zhang et al., (2019) and other literature have mentioned that populations or outliers with low density are outliers, and DBSCAN performs well in terms of outliers.

This study explores the problem of using LSTM autoencoder combined with DBSCAN model for anomaly detection under the condition of data imbalance, which can overcome the problem of requiring a large number of labels and data imbalance. By inputing normal data into the LSTM autoencoder for training, a higher detection accuracy than unsupervised learning can be obtained. The reconstruction error of the trained model is used as the model training evaluation index, and the error of each time point between the reconstruction sequence and the actual sequence is calculated. Then, the dimension of the reconstruction error was reduced by the principal components analysis (PCA). The outlier values were determined with DBSCAN according to the density of the reconstruction error. Finally, a performance evaluation was performed and compared with other literatures.

METHODOLOGY OF AUTOENCODER ANOMALY DETECTION

This study adopted the concepts proposed by Jiang et al. (2018) and Nguyen et al. (2021) to perform a multivariate time series anomaly detection for a water distribution system. The detection process consists three parts: data preprocessing, error reconstruction computation and anomaly classification.

In most cases, the data scale returned by the sensors in the manufacturing equipment will be very different, which is easy to affect the performance of the model, so the original data must be compressed and the data of different dimensions must be standardized to the same scale. In this study, the feature scales were standardized to improve the convergence speed and the accuracy of the model.

In the actual manufacturing environment, sensors of equipment usually return a considerable amount of time series data. The sliding window method can be adopted to segment the original data according the window size and sliding step size to divide data into a subseries composed of several samples. The size of the window will affect the performance of the model to a considerable extent, and a larger window can better extract the dependence between the sequence and time, but the anomaly may only account for a small part of the sequence, resulting in a decrease in the accuracy of detection, a higher Type II error, and a later time to send out the abnormal signal. Setting a smaller window will better identify the anomalous parts, but it may also ignore the dependence of sequence data and time, making it difficult for the model to learn the main features of the sample, increasing the probability of false alarms, and leading to a high Type I error.

LSTM-Autoencoder Model

The pre-processed subsequences of the original data are standardized and segmented into the model, and the model adopts semi-supervised learning, which is trained on the data in the normal state and tested with the data containing the normal and abnormal states. The model consists of two parts, a LSTM encoder and a LSTM decoder. A fully connected layer is then added after the decoder outputs the reconstruction data. The model structure is shown in Figure 1, and finally the reconstruction error is calculated with equation (1). Most studies will average the reconstruction errors of each dimension into one value, but this study retains the errors of each dimension as the basis for anomaly classification.

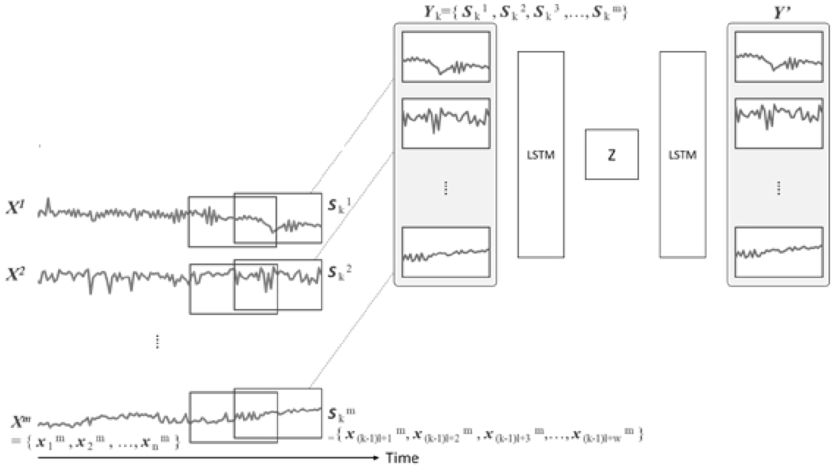
$Y_k = \{S_k^1, S_k^2, S_k^3, \ldots, S_k^m\}$
Y'
LSTM
Z
LSTM
X^1
S_k^1
X^2
S_k^2
$X^m = \{x_1^m, x_2^m, \ldots, x_n^m\}$
$S_k^m = \{x_{(k-1)l+1}^m, x_{(k-1)l+2}^m, x_{(k-1)l+3}^m, \ldots, x_{(k-1)l+w}^m\}$
Time

Figure 1: LSTM-autoencoder model.

$$e_t^m = \frac{1}{W} \sum_{i = t-w+1}^{t} \left(x_i^m - \widehat{x}_t^m\right)^2 \tag{1}$$

DBSCAN Classification Model

In the stage of the classification, the reconstruction error of m dimensions obtained by the LSTM autoencoder will be used as a feature for clustering. Since a too high dimension will affect the performance of DBSCAN grouping, PCA will be used to reduce the dimensionality of the data before the reconstruction error is input into the DBSCAN classification model. Zhou and Hou (2022) showed that the combination of PCA and DBSCAN can improve the performance of DBSCAN. DBSCAN will first determine the neighborhood radius (ε) and the minimum number of points (minPts) in the neighborhood radius, and then randomly select any core point in the data, searching for the density of data points near the core point. If the density is connected, it will be expanded to the same cluster until there are no points can be expanded. Finally, the outliers outside the cluster are treated as outliers.

Performance Evaluation

This study uses F1-score for model performance evaluation as shown in equations (2)~(5). Precision is used to evaluate the accuracy of the results. Recall-rate is used to evaluate the completeness of the results; F1-score is the harmonic average of precision and recall-rate, which is used to find a balance between precision and recall-rate, and the closer the value is to 1, the better..

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \times 100\% \tag{2}$$

$$\text{Precision} = \frac{TP}{TP + FP} \times 100\% \tag{3}$$

$$\text{Recall - rate} = \frac{TP}{TP + FN} \times 100\% \tag{4}$$

$$\text{F1 - score} = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100\% \tag{5}$$

RESULTS

This study uses a public dataset posted by Katser (2020) in Kaggle Skoltech Anomaly Benchmark (SKAB) dataset, which is a dataset of water supply equipment built in a simulation laboratory. The data set contains 34 abnormal time series of about 1000 sample points each with 13 abnormal causes, and the data is returned every second. There is no missing value in this data set, and the sensor collects and transmits back eight signals.

Firstly, the 34 data were normalized, and the sliding windows were set to 10, 20, 30, 40, 50, 60, 70, and 80 in seconds to compare the impact of window size on model performance. The encoder and decoder each had a layer and consisted of 100 neurons, using relu as the activation function, adam as the optimizer, the learning rate was 0.0001, the epoch size was 100, and the batch size was 32. In addition, a fully connected layer is added after the decoder to output the reconstructed data. Finally, the reconstruction errors of different dimensions at each time point are calculated as the input characteristics of anomaly classification model. The reconstruction error output by the LSTM autoencoder retains the original 8 dimensions of the dataset, which was reduced to 2 dimensions by PCA.

The reconstruction error was inputed into the DBSCAN model for grouping after PCA dimensionality reduction. If the point is assigned to a certain group, the point is judged to be normal, and if the point is not classified as an outlier in the DBSCAN model, it is judged to be abnormal. There are two important parameters in DBSCAN, namely: fixed neighborhood distance (ε) and minimum number of points (minPts). Since the best parameters of 34 samples are not the same, the performance of 34 samples is averaged. This study tests 8 window sizes with 50 DBSCAN parameters, a total of 400 combinations. The classification performance with ε=25 and minPts = 5 was the best as shown in Figure 2, and when the sliding window is 60, the model has the best performance.

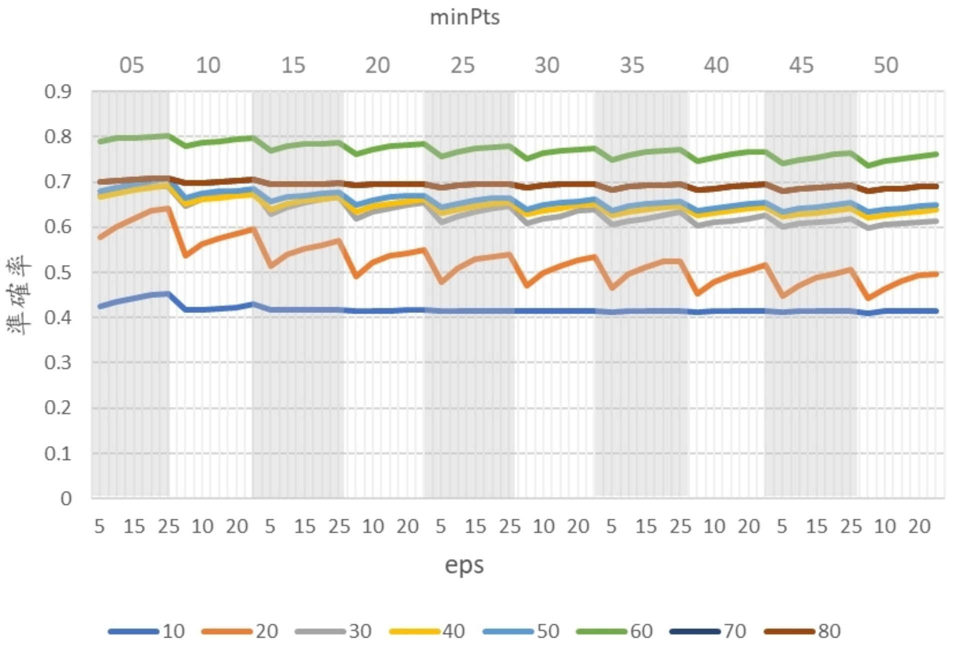

Figure 2: DBSCAN F1-score with different parameters (ε, minPts) and window sizes.

Table 1. Performance comparisons of different methods.

Source	Algorithm	F1-score
SKAB (2020)	LSTM-AE	0.5110
Moon et al. (2023)	TL-AE	0.4823
	TL-VAE	0.4821
	MAML-AE	0.6119
	MAVAE	0.7032
Proposed	LSTM-AE-DBSCAN	**0.8025**

Each sample had a more suitable sliding window, and most samples had better performance when the sliding window was 60, and the overall performance was also the best when the sliding window was 60, with an F1-score of 0.803. The F1-score classification performance was used to compare the performance of the four models(TL-AE, TL-VAE, MAML-AE, MAVAE) proposed by Moon et al. (2023) and the LSTM autoencoders proposed in the study. As shown in Table 1, our model outperform the other models.

CONCLUSION

In this study, we used the SKAB water distribution system dataset to verify the anomaly detection model of the proposed LSTM autoencoder combined with DBSACAN, and compared the impact of different sliding windows on the performance of the model. Based on the LSTM autoencoder, the enhanced DBSCAN classification model can avoid the difficulty of setting the threshold. Compared with previous studies, the overall performance of this study was better than that of previous studies.

REFERENCES

Çelik, M., Dadaşer-Çelik, F., & Dokuz, A. Ş. (2011, 15–18 June 2011). Anomaly detection in temperature data using DBSCAN algorithm. 2011 International Symposium on Innovations in Intelligent Systems and Applications.

de Pater, I., & Mitici, M. (2023). Developing health indicators and RUL prognostics for systems with few failure instances and varying operating conditions using a LSTM autoencoder. *Engineering Applications of Artificial Intelligence*, *117*, 105582. https://doi.org/10.1016/j.engappai.2022.105582

Garg, S., Kaur, K., Batra, S., Kaddoum, G., Kumar, N., & Boukerche, A. (2020). A multi-stage anomaly detection scheme for augmenting the security in IoT-enabled applications. *Future Generation Computer Systems*, *104*, 105–118. https://doi.org/10.1016/j.future.2019.09.038

Iurii D. Katser, V. O. K. (2020). *"Skoltech Anomaly Benchmark (SKAB)." Kaggle*.

Jiang, G., Xie, P., He, H., & Yan, J. (2018). Wind Turbine Fault Detection Using a Denoising Autoencoder With Temporal Information. *IEEE/ASME Transactions on Mechatronics*, *23*(1), 89–100. https://doi.org/10.1109/TMECH.2017.2759301

Li, G., & Hu, Y. (2018). Improved sensor fault detection, diagnosis and estimation for screw chillers using density-based clustering and principal component analysis. *Energy and Buildings*, *173*, 502–515. https://doi.org/10.1016/j.enbuild.2018.05.025

Luo, B., Wang, H., Liu, H., Li, B., & Peng, F. (2019). Early Fault Detection of Machine Tools Based on Deep Learning and Dynamic Identification. *IEEE Transactions on Industrial Electronics*, *66*(1), 509–518. https://doi.org/10.1109/TIE.2018.2807414

Mao, W., Ding, L., Liu, Y., Afshari, S. S., & Liang, X. (2022). A new deep domain adaptation method with joint adversarial training for online detection of bearing early fault. *ISA Transactions*, *122*, 444–458. https://doi.org/10.1016/j.isatra.2021.04.026

Moon, J., Noh, Y., Jung, S., Lee, J., & Hwang, E. (2023). Anomaly detection using a model-agnostic meta-learning-based variational auto-encoder for facility management. *Journal of Building Engineering*, *68*, 106099. https://doi.org/10.1016/j.jobe.2023.106099

Provotar, O. I., Linder, Y. M., & Veres, M. M. (2019, 18–20 Dec. 2019). Unsupervised Anomaly Detection in Time Series Using LSTM-Based Autoencoders. 2019 IEEE International Conference on Advanced Trends in Information Theory (ATIT).

Ranjith, R., Athanesious, J. J., & Vaidehi, V. (2015, 15–17 Dec. 2015). Anomaly detection using DBSCAN clustering technique for traffic video surveillance. 2015 Seventh International Conference on Advanced Computing (ICoAC).

Singh, H. V., Girdhar, A., & Dahiya, S. (2022, 25–27 May 2022). A Literature survey based on DBSCAN algorithms. 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS).

Xu, X., & Yoneda, M. (2021). Multitask Air-Quality Prediction Based on LSTM-Autoencoder Model. *IEEE Transactions on Cybernetics*, *51*(5), 2577–2586. https://doi.org/10.1109/TCYB.2019.2945999

Zhang, S., Xiao, K., Carranza, E. J. M., Yang, F., & Zhao, Z. (2019). Integration of auto-encoder network with density-based spatial clustering for geochemical anomaly detection for mineral exploration. *Computers & Geosciences*, *130*, 43–56. https://doi.org/10.1016/j.cageo.2019.05.011

Zhou, W., & Hou, J. (2022). Implementation of fault isolation for molten salt reactor using PCA and contribution analysis. *Annals of Nuclear Energy*, *173*, 109138. https://doi.org/10.1016/j.anucene.2022.109138