Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 3, No. 3, 2025

Research on Artificial Intelligence Driven Anomaly Detection Model for Big Data

Zhengyang Li [a, *], Fengrui Zhang [b]

[1] DigiPen Institute of Technology, Redmond, Washington, USA
[2] Worcester Polytechnic Institute, Worcester, MA, USA.
[a] zhengyang.li@digipen.edu, [b] zhangfengrui95@gmail.com

Abstract: Facing the high-dimensionality, heterogeneity and temporal complexity of anomaly detection in big data environment, an intelligent detection model integrating graph neural network, self-encoder and attention mechanism is designed. The model structure is equipped with multimodal feature encoding capability and online adaptive mechanism, which improves the recognition performance of rare anomalies and structural mutations. Experiments based on the KDDCup99 and NSL-KDD datasets demonstrate that the model outperforms multiple comparative methods in terms of accuracy and robustness, and shows good practicality and scalability.

Keywords: Big data; Anomaly detection; Deep learning models.

1. Introduction

In the big data environment, anomaly data is often accompanied by system risks, business anomalies or potential attack behaviors, and its detection and identification are crucial to guarantee the stability, security and reliability of the data system. Traditional anomaly detection methods have shown performance bottlenecks when dealing with high-dimensional, large-scale, and dynamically changing data, making it difficult to meet the real-time and accuracy requirements of modern data systems. Artificial intelligence technology, especially the rise of deep learning, reinforcement learning and self-supervised learning, provides new solution ideas and method support for anomaly detection, and has become one of the current research hotspots in this field.

2. Analysis of Big Data Anomaly Detection Problems

2.1. Application Challenges of Anomaly Detection

Anomaly detection has key value in several high-risk areas, including financial fraud prevention and control, network security early warning, industrial equipment maintenance and intelligent medical analysis. In practice, the system needs to identify a very small number of hidden, highly variable anomalous behaviors, which puts high demands on the detection model. Common challenges include: the scarcity of anomaly samples, which makes supervised learning difficult to train effectively; the complexity of anomaly types, which may be manifested as structural mutation, behavioral drift, etc., which is difficult to model in a unified way; the massive data mixed with a large amount of noise information and pseudo anomalies, which increases the risk of false alarms; the non-linear dependence between features in the high-dimensional feature space, which makes it difficult for the traditional algorithms to extract the stable anomaly feature representations. With the diversification of data sources and the improvement of real-time business requirements, anomaly detection requires not only high accuracy, but also the ability to recognize the dynamics of multimodal and non-smooth data.

2.2. Detection Difficulties in Big Data Environment

In the big data context, the anomaly detection task is transformed from static analysis to dynamic, online, multi-source learning, and the technical complexity rises significantly. First, the data scale grows exponentially, and traditional algorithms are difficult to support terabyte-level data processing in terms of computational and memory resources; second, the data update frequency is high, and the model needs to have the ability of streaming learning, otherwise it will miss the key anomaly events. In addition, there is a conceptual drift phenomenon of anomaly types, i.e., anomalous behaviors evolve over time and their statistical features are no longer stable, leading to the failure of static models. The high-dimensional sparsity and heterogeneity of data also further exacerbate the modeling difficulty, especially in multi-structured data scenarios such as text, graphs, logs, etc., which require the introduction of deep structured modeling techniques. Therefore, the key task of big data anomaly detection is to build an intelligent detection framework with timeliness, adaptivity and distribution generalization capabilities, breaking through the traditional detection paradigm based on static rules or shallow features.

3. Artificial Intelligence-based Anomaly Detection Model for Big Data

3.1. Model Design

Facing the core challenges of high-dimensional sparsity, heterogeneous structure and dynamic evolution faced by anomaly detection in big data environment, a composite architecture integrating self-encoder, graph neural network and attention mechanism is proposed. The model adopts a multi-module tandem structure, and the overall architecture includes: data access and preprocessing module, feature encoding module, anomaly detection main engine and

feedback update mechanism. The overall PyTorch as a modeling platform, combined with Spark streaming processing capabilities, for multi-source heterogeneous data to design a deployable and updatable AI-driven detection system, the structure is shown in the figure below:

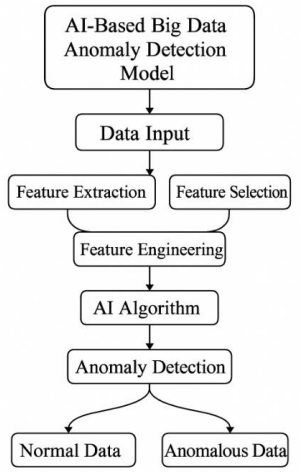

Figure 1. Flowchart of AI-based anomaly detection model for big data

The design addresses three core issues: high-dimensional time-series data, data structure complexity, and anomaly type evolution. First, in the data input layer, streaming or batch data sources are accessed and uniformly converted to tensor structure [4]. The model middle layer uses deep neural networks for end-to-end representation learning. Where temporal data (e.g., device sensor data, transaction sequences) are encoded by GRU and structural data (e.g., logs, communication graphs) are processed by GCN to preserve topological information. Subsequently, all intermediate representations are uniformly mapped to the anomaly detection engine. The detection engine is based on a Residual AutoEncoder structure (RAE), which identifies potential anomaly regions by maximizing the reconstruction error, and introduces a multi-attention mechanism to model dynamic dependencies among features. In order to improve the adaptability of the system, the model deploys an adaptive feedback module at the output side to softly update based on the prediction confidence, and at the same time builds a sample cache queue to regularly update the model parameters to realize online learning and stability control. The whole system supports micro-batch training and can be deployed in GPU cluster or edge nodes.

3.2. Core Algorithm

The core detection algorithm of the model integrates the reconstruction error-driven detection mechanism with the graph representation enhancement mechanism. The backbone network adopts a residual self-encoder structure, i.e., cross-layer residual connections are introduced into the standard AutoEncoder to enhance the information retention ability of the deep network. In the encoding stage, the input data $X \in R^{n\times d}$ is mapped by a nonlinear mapping to obtain a low-dimensional potential representation $Z \in R^{n\times k}$, and the decoder tries to reduce $\hat{X}$ and use the reconstruction error $E = \left\| X - \hat{X} \right\|_2^2$ as an anomaly metric. In order to overcome the problem of inter-node dependencies that cannot be modeled in complex graph-structured data, graph convolutional network (GCN) module is introduced in the design, which is embedded by taking the graph adjacency matrix A and node feature matrix H as inputs in the following form:

$$H^{(l+1)} = \sigma(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{(l)}W^{(l)})$$

Where $\tilde{A} = A + I$ denotes the plus self-loop adjacency matrix, $\tilde{D}$ is the degree matrix, $W^{(l)}$ is the learnable parameter, and σ is the activation function. The GCN embedded output is spliced with the encoder output for anomaly determination. The whole anomaly metric function is defined as:

$$S(x) = \alpha \cdot \left\| x - \hat{x} \right\|_2^2 + \beta \cdot \left\| f_{gcn}(x) - \hat{x}_{gcn} \right\|_2^2$$

Where α, β controls the weight of reconstruction error and graph representation error. In order to further improve the detection sensitivity, the attention mechanism is introduced in the coding layer, and a learnable weight vector is introduced to the feature dimensions, and the importance of different features is calculated by softmax, so as to dynamically adjust the anomaly discrimination criterion. The overall optimization goal of the algorithm is to minimize the reconstruction error distribution of normal samples, and to make the abnormal samples maintain a significant distance from the normal samples in the embedding space, so as to guarantee the clarity of the detection boundary.

3.3. Feature Processing Methods

Data preprocessing and feature engineering are differentiated for different data types. Structured data, such as transaction records and equipment monitoring data, are cleaned of abnormal outliers using normalization and Z-Score methods, and time series are sliced by sliding windows. Statistical features (mean, extreme deviation, kurtosis), rate-of-change features, and periodic coding are extracted within each window to form a preliminary feature tensor. Unstructured data (e.g., system logs or network messages) are first extracted by regular matching to extract key fields, and then the BERT embedding model is used to generate a context-aware vector representation [5]. For graph data (e.g., communication structures, IoT sensing graphs), graph attribute features based on node behavior frequency and neighbor variability are constructed, and topology vectors are constructed using the Node2Vec method. Instead of using artificial labels, the pseudo-labeling mechanism is used to initialize the sample categories in the model training stage: firstly, the deviation of the reconstruction error of each sample from its K-nearest-neighbor average error is calculated, and the initial "credible anomaly" candidate set is constructed, which is used to supervise the initialization training process [6-7].

4. Experimental Validation and Effect Evaluation

4.1. Experimental Setup

In order to verify the anomaly detection performance of the constructed model in a big data environment, two widely used network intrusion detection datasets, KDDCup99 and NSL-KDD, are used in the experiments, which represent the classical big data and medium-sized scenarios, respectively [8]. The experimental platform is based on NVIDIA A100

GPU server, PyTorch + DGL environment is built, and the data preprocessing uses Spark to realize parallel processing and tensorization transformation of feature engineering. The Adam optimizer is used in the training phase, the initial learning rate is set to 0.001, the batch size is 256, the number of model training rounds is 100, and the early-stop strategy is dynamically adjusted based on the F1 index of the validation set [9]. In order to comprehensively evaluate the model performance, a comparison experimental set is set up including typical methods such as Isolation Forest, AutoEncoder, GCN and LSTM-AE, and the anomaly detection task under unsupervised setting is taken as the benchmark uniformly. The evaluation indexes were chosen as Accuracy, Recall, F1 value and area under the AUC curve to comprehensively reflect the recognition ability and stability of the model. All experiments were repeated 5 times and averaged to minimize the effect of chance.

4.2. Model Performance Analysis

The experimental results show that the proposed fusion model significantly outperforms the traditional method in several key indexes, and possesses stronger abnormality discrimination ability and stability. On the KDDCup99 dataset, the model achieved an accuracy of 98.3% under the unsupervised setting, with an F1 value of 96.8%, which is significantly higher than that of other methods; it maintains a good migration robustness on NSL-KDD, and exhibits consistency in recognizing anomaly categories [10]. Table 1 summarizes the performance comparison of each method on KDDCup99:

Table 1. Performance comparison of methods on KDDCup99 dataset

Method Name	Accuracy (%)	Find all rate (%)	F1 value (%)	AUC
Isolation Forest	90.2	88.5	86.9	0.917
AutoEncoder	94.8	91.2	90.5	0.948
GCN	96.1	93.4	92.6	0.962
LSTM-AE	95.6	92.8	91.4	0.956
Fusion model (this model)	98.3	96.2	96.8	0.981

In real business scenarios, the model possesses higher anomaly response accuracy, especially in low-frequency anomalies and structural attack pattern identification, which shows strong generalization ability. This verifies the advantage of the multimodal depth structure in capturing the correlation features and dynamic evolution patterns among complex data.

5. Conclusion

The proposed detection model constructs a multilayer fusion mechanism for complex data structures and dynamic anomaly patterns, which significantly improves the system's ability to recognize high-dimensional anomaly features. It maintains the performance stability in streaming processing and unsupervised environment, and has good prospects for engineering applications. Subsequent work will further focus on the model's ability to adapt to the evolution of anomaly categories under conceptual drift, enhance the generalization and long-term effectiveness of the model by introducing cross-modal comparative learning and structural self-supervision methods, and improve the system's deployment flexibility and risk prevention and control ability in real-world scenarios.

References

[1] Jing Zhang. Research on the algorithm of environmental monitoring data processing and anomaly identification based on artificial intelligence [J]. Chinese Science and Technology Journal Database (Full Text Edition) Natural Science, 2025(1):132-135.

[2] Li Yi. Anomaly detection algorithm for power communication transmission network based on artificial intelligence [J]. Communication Power Technology, 2025, 42(2):242-245.

[3] HONGSONG CHEN, XINRUI LIU, ZIMEI TAO, ZHIHENG WANG. A research review on deep learning-based anomaly detection for timing data [J]. Information Network Security, 2025(3):364-391.

[4] Wang Y. Construction of a Clinical Trial Data Anomaly Detection and Risk Warning System based on Knowledge Graph [C]//Forum on Research and Innovation Management. 2025, 3(6).

[5] Xiang Y, Li J, Ma K. Stock Price Prediction with Bert-BiLSTM Fusion Model in Bimodal Mode [C]//Proceeding of the 2024 5th International Conference on Computer Science and Management Technology. 2024: 1219-1223.

[6] Ravula R K. Leveraging AI-Driven Anomaly Detection for Enhanced Data Quality and Regulatory Compliance in Clinical Studies [J]. Journal of Computer Science and Technology Studies, 2025, 7(2): 240-248.

[7] Gancheva V. Software Anomaly Detection Method Based on Artificial Neural Network [C]//2024 IEEE International Conference on e-Business Engineering (ICEBE). IEEE, 2024: 272-277.

[8] Yuhertiana I, Amin A H. Artificial Intelligence Driven Approaches for Financial Fraud Detection: A Systematic Literature Review [J]. KnE Social Sciences, 2024: 448–468-448–468.

[9] Gancheva V. Software Anomaly Detection Method Based on Artificial Neural Network [C]//2024 IEEE International Conference on e-Business Engineering (ICEBE). IEEE, 2024: 272-277.

[10] Jung J, Park S, Kim H, et al. Artificial intelligence-driven video indexing for rapid surveillance footage summarization and review [C]//Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence. 2024: 8687-8690.