Sequence Anomaly Detection based on Diffusion Model

장지원 1, 조인휘 2
1,2 한양대학교 컴퓨터소프트웨어학과

zhiyuan15@hanyang.ac.kr, iwjoe@hanyang.ac.kr

확산 모델 기반 시퀀스 이상 탐지

Zhiyuan Zhang[1], Inwhee Joe[2]
[1,2]Dept. of Computer Science, Hanyang University

Abstract

Sequence data plays an important role in the field of intelligence, especially for industrial control, traffic control and other aspects. Finding abnormal parts in sequence data has long been an application field of AI technology. In this paper, we propose an anomaly detection method for sequence data using a diffusion model. The diffusion model has two major advantages: interpretability derived from rigorous mathematical derivation and unrestricted selection of backbone models. This method uses the diffusion model to predict and reconstruct the sequence data, and then detects the abnormal part by comparing with the real data. This paper successfully verifies the feasibility of the diffusion model in the field of anomaly detection. We use the combination of MLP and diffusion model to generate data and compare the generated data with real data to detect anomalous points.

1. Introduction

Anomaly detection in sequence data is an important problem in many fields, such as finance, healthcare, and cybersecurity, where the ability to detect unusual patterns can provide valuable insights and help prevent potential problems. For the anomaly detection of sequence data, many scholars have proposed various methods. The isolation forest method, proposed by Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou, utilizes the principle that anomalies are isolated more quickly than normal data points through a random partitioning approach, and has been shown to be effective for anomaly detection in various applications [1]. Li et al. proposed a method using GANs to find anomalous parts in sequence data by comparing real and generated values [2]. Inspired by this approach, we propose a diffusion model-based anomaly detection method.

The emergence of OpenAI's image-text pair generation model DALLE2 made the diffusion model [3] suddenly explode. Unlike GAN, which is also a generative model, the diffusion model can generate different styles for the same object when generating pictures.

Diffusion models have two major advantages when dealing with anomaly detection tasks on sequence data. The first point is that the prediction model in the diffusion model can be any model that meets the task requirements. This is very useful for anomaly detection tasks with complex and changeable task environments. The second point is that the diffusion model has a complete mathematical reasoning process, which has a very high interpretability. In this paper we investigate the feasibility of diffusion models for anomaly detection tasks.

2. Method

The diffusion model generally consists of two parts, the diffusion process of forward propagation and the generation process of back propagation. Let's take the S-shaped curve as an example. Figure 1 shows the process of forward diffusion. We ended up turning the curve into a completely random distribution by adding noise repeatedly. Figure 2 shows the reverse generation process. From the Gaussian distribution, we can finally get a shape that approximates the preset curve.

To realize the function of anomaly detection, we adopt the method of numerical comparison. First, we'll set a threshold, subtract the actual value from the predicted value, and compare it to that threshold. If the value obtained after subtraction is greater than the threshold, we consider the part where the value is located to be an outlier. It is worth noting that the data we use during training should not include outliers.

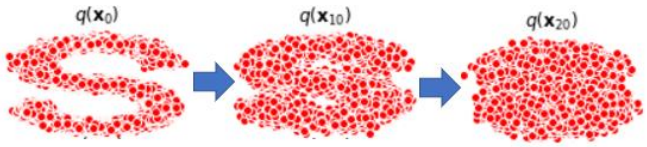

Figure 1. Forward diffusion process

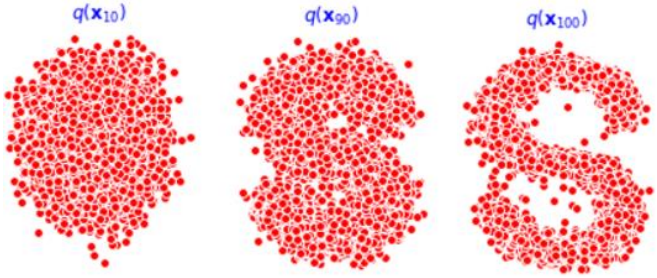

Figure 2. Generation process

3. Experiments

In this paper, instead of using the complete external dataset, we use the generated sine function data as training data. The data used in the training process does not contain outliers. Figure 3 shows part of the training dataset.

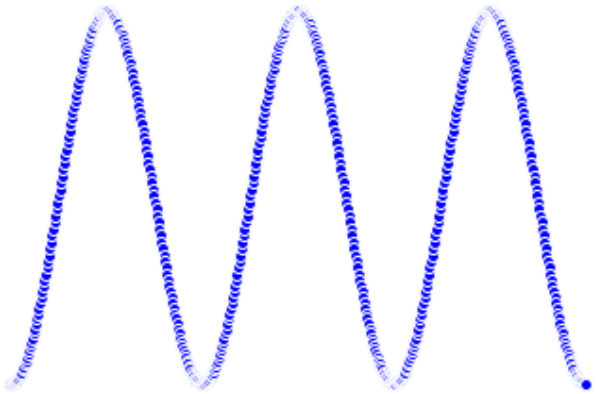

Figure 3. Dataset

As the number of iterations increases, the generative ability of the model becomes more and more accurate. Figure 4, Figure 5, and Figure 6 show the sequence data generated by the model at 100epoch, 1500epoch and 8000epoch, respectively. We can clearly find that the generation ability of the model continues to increase, and the generated data is getting closer and closer to the original data.

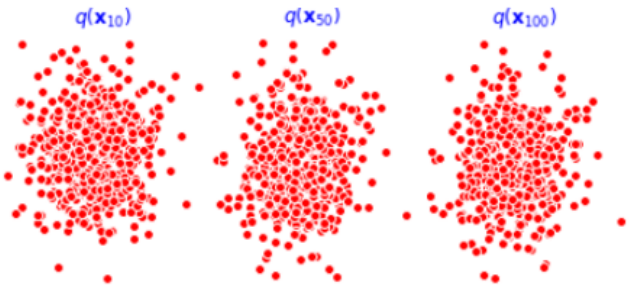

Figure 4. Data generated at 100 epochs.

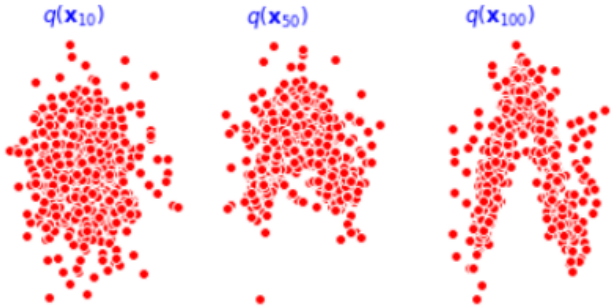

Figure 5. Data generated at 1500 epochs.

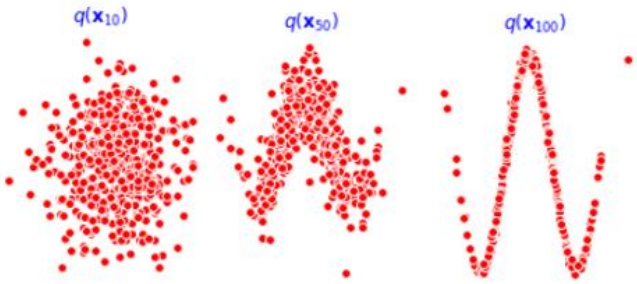

Figure 6. Data generated at 8000 epochs.

Number of Anomaly points	Accuracy
1	100%
10	90%
100	88%
500	85%

Table 1. Number of anomaly points and Accuracy

In table 1., We set different numbers of anomalous points in the experiment. As the number of anomalous points increases, the accuracy of the model decreases. However, overall, the model still performs well.

After the model generates data that is close to the original data, exception handling will also become relatively easy. We put some outliers in the original data to test whether the method of comparing the generated data with the test data can identify the outliers.

4. Conclusion & future work

This paper explores the application of a new generative

model diffusion model to the task of anomaly detection on sequence data. We use sinusoidal function data as training data and define outliers by comparing the difference between the data generated by the model and the actual data.

Through experiments we found that the diffusion model can complete and realize this task. At the same time, we also found that the diffusion model has generated data that is closer to the original data for shorter sequence data.

For future work, we can explore the diffusion model for long-sequence data processing or use the sliding window method to improve the model effect.

Reference

[1] Liu F T, Ting K M, Zhou Z H. Isolation forest[C]//2008 eighth ieee international conference on data mining. IEEE, 2008: 413-422.

[2] Li D, Chen D, Jin B, et al. MAD-GAN: Multivariate anomaly detection for time series data with generative adversarial networks[C]//Artificial Neural Networks and Machine Learning–ICANN 2019: Text and Time Series: 28th International Conference on Artificial Neural Networks, Munich, Germany, September 17–19, 2019, Proceedings, Part IV. Cham: Springer International Publishing, 2019: 703-716.

[3] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[J]. Advances in Neural Information Processing Systems, 2020, 33: 6840-6851.