Time series data anomaly detection based on LSTM-GAN

Xiaofei Pan*

School of Control and Computer Engineering, North China Electric Power University, Baoding 071003, Hebei Province, China

* Corresponding author: Xiaofei Pan (Email: px448265572@163.com).

Abstract: With the improvement of modern technology, a large number of time series data have been produced. The anomaly detection of time series data can provide relevant information for key situations faced by various fields. This paper proposed an unsupervised temporal anomaly detection method based on generation countermeasure network. In this model, Wasserstein distance is used instead of the original measurement method, and LSTM is used as the basic network of GAN. The model uses the reconstruction loss of the generator and the loss of the discriminator to define the anomaly function to judge the anomaly. This paper uses real world time series data sets involving various fields to evaluate the model. Experiments show that the model is effective in anomaly detection of time series data.

Keywords: Time series data; Anomaly detection; GAN.

1. Introduction

Exceptions can be defined as abnormal patterns that do not conform to expected behavior. [1] With the rapid growth of time series data, abnormal data may lead to serious failures. Anomaly detection (AD) refers to automatic identification of abnormal phenomena mixed with a large number of normal data. [2] Therefore, anomaly detection can provide an opportunity to take action and solve problems before potential problems cause disasters.

Traditionally, various statistical methods have been proposed to improve the threshold, such as SPC [3]. If the monitoring data calculated on an instance exceeds the control limit, it will be identified as an exception. Such methods require a lot of human knowledge to set the prior assumptions of the model [4].

Later, some unsupervised machine learning methods were proposed. A common method is to divide time series data into subsequences with a certain length, and use clustering method to find outliers. The other is to learn a model to predict or reconstruct time series data, and compare the actual value with the predicted value or the reconstructed value. A high prediction error or reconstruction error calculated indicates the existence of anomalies [5].

With the rapid development of artificial neural networks in various fields, more and more methods based on deep learning have emerged. GAN is widely used in many fields, such as image generation, data balance, image recognition and video processing. [6] GAN can learn the real data distribution and generate false data similar to the real data. When the input is an abnormal sample, the sample generated by GAN will be different from the original real data. When the difference is greater than the threshold, it will be judged as abnormal.

In this paper, a temporal data anomaly detection model called LSTM-GAN is proposed. The structure of the rest of this paper is as follows: In section 2, the relevant technologies used are introduced, in section 3, the model structure proposed in this paper is introduced, in section 4, the relevant experiments conducted with the model proposed in this paper are introduced, and the last part is summarized.

2. Related technologies

The generative adversarial network is based on the scenario of game theory, [7] in which two participants compete with each other to achieve the goal of Nash equilibrium. The GAN network flow chart is as follows:

![GAN network flow chart](image)

GAN is composed of a generator G and a discriminator D. The random variable z from potential space Z is input into generator G to learn the distribution on real samples by mapping G(z). Generator G needs to generate generated samples that are as similar as possible to the distribution of real samples. The loss function of generator G is:

\[
L_G = E_{z \sim p_{z}}[\log(1 - D(G(z)))]
\]

Input the real sample x or generated sample into discriminator D, and its training goal is to correctly distinguish between the real sample and generated sample.

Generator G and discriminator D play games through min-max function.

After training, generator (G) has the ability to generate false data Discriminator (D) has the ability to distinguish between real data and false data.

Assuming \( p_{data(x)} \) and \( p_{data(x)} \) are known, the optimal discriminator function is:
The discriminator needs to distinguish between real samples and false samples. Since it is necessary to fit Wasserstein distance, the objective function is:

$$
\min_G \max_D V(C,G) = E_{\gamma \sim \text{mix}}(C(x)) + E_{\gamma \sim \text{mix}}(1-D(G(z)))
$$

4. Model Evaluation

In this paper, real world datasets from NAB dataset are used in the experiment. Since F1 can give consideration to both the accuracy and recall of the model, it serves as the main indicator of the evaluation model. The experimental results are as follows:

<table>
<thead>
<tr>
<th>Methods</th>
<th>Art</th>
<th>AdEx</th>
<th>AWS</th>
<th>Traf</th>
<th>Tweets</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM</td>
<td>0.375</td>
<td>0.538</td>
<td>0.474</td>
<td>0.643</td>
<td>0.543</td>
<td>0.509</td>
</tr>
<tr>
<td>Arima</td>
<td>0.353</td>
<td>0.583</td>
<td>0.518</td>
<td>0.571</td>
<td>0.567</td>
<td>0.524</td>
</tr>
<tr>
<td>LSTM-GAN</td>
<td>0.545</td>
<td>0.632</td>
<td>0.676</td>
<td>0.684</td>
<td>0.678</td>
<td>0.643</td>
</tr>
</tbody>
</table>

Bold font is the highest value. It can be seen from the table that the model proposed in this paper performs best on each dataset.

5. Conclusion

This paper proposes LSTM-GAN model for anomaly detection of time series data. This model uses LSTM as the basic network of generator and discriminator, and uses Wasserstein distance to replace the original measurement method. In order to verify the performance of the model, the model is tested on several real-world datasets, and the model proposed in this paper has good anomaly detection results.

References
