An Anomaly Pattern Detection Method Based on Spatial Density of Electric Power Sensor Data

Weijiang Lu¹, Shuya Lei¹*, Weiwei Liu¹ and Bin Yu¹

¹Artificial Intelligence on Electric Power System State Grid Corporation Joint Laboratory, Global Energy Interconnection Research Institute Co. Ltd, Beijing, China.

*Corresponding Author Email: leishuya@geiri.sgcc.com.cn

Abstract. Electric power sensor data is a typical time-series data. Their anomalies are not only abnormal points, but also continuous data fragments named anomaly patterns. Based on the analysis of the abnormal patterns of electric power sensor data, we propose an abnormal pattern detection method to handle the anomaly of spatial density of power sensor data. The proposed method trying to discover abnormal density distribution in the super high-dimensional space. It takes the spatial distribution of the sensor data to build a density model. Moreover, density value of each sensor is computed by matching original sensor data with the density intervals in the model. Furthermore, anomalies are detected according to the density values. In order to verify effectiveness of the method, exclusive experiments are conducted on the real data of power plants. Experimental results show that the proposed method has high recall rate, low false alarm rate and high accuracy while low cost of time when detecting abnormal data patterns.

1. Introduction

The fast development of electric power Internet of Things provided amount of data for enterprises. Based on these data, many data analysis tasks are carried out to provide more accurate services [1]. However, the electric power Internet of Things environment is extremely complex. Specifically, with the improvement of power system construction, the scale of measuring points is gradually expanded. Large number of sensors are deployed and also lead to various problems, such as equipment failure, signal interference, abnormal transmission and so on [2].

As a result, there are many anomalous data in these sensor data. Nevertheless, if these anomalies are not addressed and analyzed effectively, they will bring some potential risks for further tasks. On one hand, abnormal data will affect the accuracy of data analysis results and lead to wrong decisions. On the other hand, if the anomalies in the sensor data can be identified as early as possible, some further financial loss could be avoided. Therefore, anomaly detection of electric power sensor data is particularly important.

Hawkins [3] defines anomalies as distinctive data in the data set, which makes people suspect that these data are not generated by random deviations, but by completely different mechanisms. Traditional anomaly detection is mostly aimed at outliers detection and many approaches are implemented based on statistics, distance, density, and clustering.

However, sensor data are a kind of typical time-series data, which own not only outlier anomalies, but also have time series segment anomalies named pattern anomalies. In this paper, pattern anomaly is defined as the trend of data segment is obviously different from other similar data segments.

Usually, the industry deployed multiple types of sensor devices to monitor the same physical device. For instance, various types of sensor data are combined to describe the working conditions of...
industrial equipment or environment [4]. In this circumstance, sensor data own the characteristics of high dimensional and timely changing. In contrast, traditional anomaly detection algorithms are focused on outlier detection, which can effectively detect outliers and not suitable for anomaly pattern detection. However, most of the anomalies shown in industrial sensor data are pattern anomalies. Hence, these algorithms often fail to achieve ideal results for industrial sensor data. What is more, due to the high-dimensional of industrial sensor data, efficiency is another problem.

This paper proposed the abnormal pattern detection method based on the power sensor data considering its high-dimensional and other characteristics. We aim to improve the efficiency of anomaly detection algorithm while keep its accuracy.

2. Characteristic Analysis of Power Sensor Data

Electric power sensor data refer to the power data collected by sensor devices, which are often used to continuously monitor the power system. They own many special properties [5-7] which are listed below:

First, electric power sensor data is continuously generated by sensors. Generally, sensor data are collected according to a fixed frequency. Time is one of the essential attributes and the most basic characteristics of electric power sensor data.

Indeed, spatial-temporal correlation [8]. Sensors are usually used to perceive the information of the physical world, which is manifested in temporal and spatial correlation.

Moreover, data similarity [9]. Based on the spatial-temporal correlation characteristics of power sensor data, there is similarity between sensor data collected by similar sensors located in similar time and space. In addition, if the monitored objects of the sensors have similar behaviors, the sensor data used to describe the behaviors should also be similar. Thus, it can be viewed that there is data similarity between the same kind of power sensor data under similar time, space or behavior conditions.

Fourth, high-dimensionality [10]. One single sensor data cannot describe the complex physical world. Therefore, it is often necessary to combine various types of sensor data to describe the state of a physical entity to form a high-dimensional sensor data. Especially in the power industry, a large number of dimensional attributes are often needed to effectively reflect the meaning behind it. For example, the electric power sensor data that monitor the working conditions of thermal generators own tens of dimensions.

Among the above features, high-dimensionality has received less attention, consequently this paper focuses on high-dimensionality and conducts experiments on data pattern anomaly detection.

As shown in Figure 1, the left graph shows the daily global radiation data collected by the sensor of photovoltaic power plant. The horizontal axis is time, and the vertical axis is daily global radiation. The right graph of Figure 1 shows the power generation of photovoltaic power plant on the same day. It can be seen from the figure that with the increase of light radiation, the detected power generation also gradually increases. Similarly, if the day is cloudy and the sunshine is unstable, the power generation data will also change with the changes in the sunshine radiation. However, due to the high
dimension of power sensor data, there will be a variety of anomalies coexisting in the super-dimensional space [11].

Based on the above analysis, it is found that the numerical density distribution of the data in the super-dimensional space will be far from the normal data. In other words, abnormal data will be in a low-density area in the super-dimensional space. We utilize the spatial density characteristics of power sensor data to identify abnormal patterns, and propose an abnormal pattern detection method based on the spatial density of power sensor data.

3. Anomaly Pattern Detection Method Based on Spatial Density of Power Sensor Data

In order to ensure the full coverage of abnormal situations as widely as possible, this paper proposes an anomaly pattern detection method based on the spatial density of power sensor data. This method obtains the density model of the sample set through the data modeling. Additionally, it matches the density coefficient of each sensor with the density model, and uses the density coefficient to judge whether it is abnormal or not.

3.1. Overview of Method

In the actual production process of power system, there are different anomalies in multi-dimensions. Particularly, the value of one-dimensional attribute shows abnormal change trend, while the values in the other dimension attribute show abnormal change amplitude. As a result, it is difficult to describe the overall abnormality clearly only through frequency domain features or motion features. Moreover, through observation, it is found that the path points of abnormal sensors data curves are quite different from those of normal sensors. Specifically, the path points of normal sensors tend to be more concentrated.

As shown in Figure 2, normal data tend to be concentrated in an interval in the spatial coordinate system, which makes the data density in this interval relatively high, while abnormal data are often far away from the interval of normal data, which makes it in a low-density area spatially. It can be seen that the path points of abnormal data are often different. Therefore, this paper takes advantage of this characteristic, uses the idea of space coordinate system, and puts forward an anomaly pattern detection method based on spatial density.

The abnormal pattern detection method based on spatial density mainly uses the spatial distribution of data to detect abnormal patterns. The method is mainly divided into two stages, namely data modeling and matching judgment, as shown in Figure 3.
Data modeling is mainly to divide all data into equal parts in a certain period of time, and calculate how many sample points in each attribute value segment of the whole data segment at each time point, so as to obtain an overall data density model for the next matching judgment.

Matching judgment is mainly to match the value of each sensor at each time point with the corresponding point in the data density model, obtain the density value of this point, and finally count the density value to make abnormal judgment.

3.2. Data Modeling

3.2.1. Feature analysis. Density-based anomaly detection algorithms [12] are often closely related to distance-based anomaly detection algorithms. By calculating the distance between sample points and cluster center, we can judge whether sample points belong to this cluster. And the number of sample points in the cluster is used to calculate the density of each point, so as to determine whether it is abnormal or not. However, this method can only show good detection performance when detecting outliers, and is not competent for detecting time series data such as power sensor data in this paper. If the data at each time point are detected in the form of outlier detection, when the amount of data is large enough, it will inevitably incur a lot of time overhead. Therefore, the traditional density-based anomaly detection algorithm cannot reach the ideal effect.

In addition, due to the high-dimensional characteristics of power sensor data, the attribute dimensions contained in the data are often as high as tens of dimensions. In this way, if the density of data is obtained by distance calculation, it will take a lot of time, and the calculation efficiency cannot be applied to the actual production situation.

Therefore, the method makes full use of the space coordinate system of the super-dimensional space to abstract all dimensions of the power sensor data to the super-dimensional space. One attribute dimension represents a spatial dimension, so that when the dimension of power sensor data is X, the hyperspace is abstracted into X-dimensional space. At the same time, the coordinate origin is the zero point of each attribute value in this space, and the coordinate system constructed is the hyper-dimensional space coordinate system. Then, let the sample data construct a density model in the super-

![Flowchart](image-url)
dimensional space, use the number of sample points in each segment as the density value of this segment by simple division, and obtain the density value by matching the sample points of the sensor to each interval. Then use this density value to judge the abnormal conditions.

3.2.2. Equal division modeling. Equal division modeling means that the sample points of the entire data segment at each time point are divided equally in each attribute value segment according to the value range, and then matched with the existing data to obtain the density model at each time point, and integrate them in chronological order. Finally, the overall data density model is obtained and used for the next matching judgment.

As shown in Figure 4, the equal division modeling draws on the idea of a multidimensional spatial geometric coordinate system. The values of the same time attribute of each sensor are divided into equal parts according to the maximum and minimum values. The size of the division determines the range of each unit lattice. The more the size, the more equal parts, the smaller the unit lattice range, the higher the accuracy, but the speed of modeling and matching will decrease. The method considers the balance of time and accuracy, chooses to model all attribute values in decimal from, and then calculates the density of all sensor numbers \{T1, T2, ..., Tn\} according to time and attributes. Specifically, if the value of a sensor at a certain point in time is within the value range of one of the unit cells, the density value of this unit cell is increased by one. After matching all sensor data, a data model with density values is obtained.

![Figure 4. Equal division modeling process.](image)

Figure 5. The specific process of data modeling.
As shown in Figure 5, in order to better describe the specific operation process of data modeling, the figure simulates the data of a certain dimension attribute, the horizontal axis is the collection time, and the vertical axis is the specific collection value. The method divides the collected value into ten equal parts according to the maximum and minimum values. Each interval represents a modeling unit. When the value collected at a certain time point is within a certain interval, this interval adds 1 to the density parameter represented, where the minimum and maximum values belong to the first and tenth intervals respectively. When the collected value is greater than or equal to the smaller segmentation point and less than the larger segmentation point, the interval adds 1 to the density parameter. When the amount of attributes other than the time attribute is greater than 2, the entire model will be built on hyperspace. After modeling all sensor data, the entire density model is built.

### 3.3. Matching Judgment

Matching judgment uses the data density model obtained in the data modeling stage to match the sample value of each sensor, so as to use the corresponding interval in the data density model to obtain the density value of the interval, and finally count the total density value of each sensor, and determine abnormal.

![Figure 6. The flow of matching judgment.](image)

As shown in Figure 6, the matching judgment stage uses the values of each sensor \{DL1, DL2, ..., DLtimemax\} to match the model \{Dm1, Dm2, ..., Dmtimemax\} completed in the data modeling stage with the density value, and the sensor’s attribute value DLi is compared with each value range of Dmi in the density model, the attribute value is matched according to the value range and the density value in the cell is obtained, and the density value is regarded as the sensor in a certain The density value at the time point, by matching the density model of each time point, the density value dvi of each attribute of the sensor at each time point can finally be obtained. After that, add up all the density values of the sensor to obtain the total density value Dvi of the sensor, and then determine whether it is abnormal according to whether the total density value of the sensor is greater than the overall average density value of all sensors. The calculation method of the overall density average is shown in Eq. (1).

\[
D_{\text{avg}} = \frac{D_{v1} + ... + D_{vn}}{n}
\]  

(1)
4. Experiment and Analysis

4.1. Experimental Environment and Dataset

In order to evaluate the effectiveness of the method, a dataset of photovoltaic power plants was selected for experimental verification. The main sources of the data set are multiple sensors on the photovoltaic arrays in the photovoltaic power plants. The entire photovoltaic array contains 54 photovoltaic modules. The sampling period of these sensors is from January 1, 2016 to December 31, 2016. And the sampling frequency is 1 minute. Each module collects 8 types of values, including total radiation value, direct radiation value, and scattered radiation value, component temperature, ambient temperature, air pressure, relative humidity and power generation. Because of the photovoltaic array was deployed in the same plant, it can be considered that the values obtained by the same attribute in each module are similar. Since the staff of the photovoltaic power plant will record the time of abnormality in the photovoltaic array, the specific serial number of the abnormal photovoltaic module and the cause of the abnormality. Therefore, according to the staff’s records of the abnormal conditions of the equipment, specific abnormal mode detection experiments are conducted for different abnormal patterns.

The data on July 12, 2016 was selected as the experimental data. The abnormal situation record of the day showed that a maintenance inspection was carried out at 5 o’clock on the same day and it was found that one of the photovoltaic modules in the photovoltaic array had equipment failure, resulting in the component temperature and direct radiation collected by it were abnormal. By observing the data, it was found that the data collected by this module began to appear abnormal from 3 pm, so the experiment selected the data of this time period for the experiment. The data set situation is shown in Table 1.

| July 12 | 53 | 1 | 6360 | 120 |
|---------|----|---|------|-----|
| Number of normal modules | Number of abnormal modules | Number of normal samples | Number of abnormal samples |

Observing the performance of the data set in the component temperature from 3 o’clock to 5 o’clock, the downward trend of the normal data sample is obviously different from the downward trend of the abnormal data sample, and the specific manifestation is that the numerical value decreases at a different speed. As for the performance of this data set in direct radiation, the overall trend of normal data samples first rose and then fell, while the overall trend of abnormal data samples showed a gradual decline. As shown in Figure 7.

![Figure 7](image)

Figure 7. Anomaly detection performance in the dataset.

In order to increase the richness of the experiment, this paper simulated similar data density distribution abnormalities, and fully tested the effectiveness of the method through experiments with different abnormal proportions. This paper simulates the experimental data of 5% abnormal sample size and 10% abnormal sample size, named "Data Set 1-1" and "Data Set 1-2" respectively.
5. Experimental Results and Analysis

5.1. Evaluation Metrics \[13\].

The evaluation indicators used in the experiment include recall rate, false alarm rate and accuracy rate. In the process of anomaly detection, experimental samples are usually divided into positive and negative types. The positive class is the minority class with few examples but distinct characteristics. The negative class is the remaining majority class.

a) Recall: For the original sample, it indicates the proportion of the positive class in the sample that is predicted correctly, that is, the proportion of the actual abnormality that is correctly detected. There are two possibilities for the total. One is to predict the original positive class as a positive class (TP), and the other is to predict the original positive class as a negative class (FN).

b) False Alarm Rate/False Positive Rate (FPR): The proportion of negative classes in the sample is predicted to be positive, that is, the proportion of normal samples that are wrongly judged as abnormal. The total is the normal sample in the data set.

c) Accuracy: The ratio of correctly predicted in the sample set. Generally speaking, the higher the accuracy, the better the anomaly detection effect, which can intuitively evaluate the effectiveness of the algorithm. The total is all samples in the data set.

5.2. Performance Experiment and Analysis.

The method was compared with the real data set, simulated data set 1-1 and simulated data set 1-2 to fully reflect the detection effect of the method under corresponding abnormal conditions. The experimental results are shown in Figure 8, Figure 9 and Figure 10.

![Figure 8](image1.png)

**Figure 8.** The recall rate of experimental results.

As shown in Figure 8, the experimental results of the method on real data and simulated data show that its recall rate is basically around 80%, indicating that the method can detect most of abnormal patterns.

![Figure 9](image2.png)

**Figure 9.** The false alarm rate of experimental results.
As shown in Figure 9, the false alarm rate remains between 1%-2%, and the labeling method is unlikely to misjudge normal power sensing equipment.

![Accuracy](image)

**Figure 10.** The accuracy of experimental results.

As shown in Figure 10, the accuracy rate remains above 97%, and the labeling method can accurately judge the abnormal conditions of most power sensing devices.

6. Conclusion
In order to better detect the pattern anomalies in the power perception data, this paper proposes an abnormal pattern detection method based on the analysis of the characteristics of the power perception data. The main idea of the approach is based on the spatial density of the power perception data. The data modeling can obtain the density model of the super-dimensional space, and the matching can obtain the density value of each sensor, so as to realize anomaly detection. In order to verify the effectiveness of the method, this paper conducts experiments on real power plant data and simulated data sets. The experimental results show that the method has a high recall rate, low false alarm rate and high accuracy rate when detecting samples with abnormal numerical density distribution. This method can effectively detect the abnormal situation of the power sensing equipment, and at the same time, it can improve the efficiency of maintenance and reduce the loss caused by the failure through real-time monitoring of the abnormal situation.

7. Acknowledgment
1. This work is supported by Research on data governance and knowledge mining technology of power IOT based on Artificial Intelligence. 2. Bin Yu were internship student in Global Energy Interconnection Research Institute co., limited during the year 2020.

8. References
[1] Yuan Wang, Ye Tao, Yingming Jiang et al. Real-time processing system for time series big data in smart power grid[J]. *Journal of Computer Applications*, 2015, 35(s2):88-92.
[2] Ye Tian, Cheng Zhang, Xinru Mao, Ji Liu, Research on Abnormal Behavior of Power ConsumptionBased on BP Neural Network with PCA, *Journal of Chongqing University of Technology (Natural Science)*. 2017, 31(8):125-133
[3] Atkinson A C, Hawkins D M. Identification of Outliers[J]. *Biometrics*, 1981,37(4):860
[4] Karkouch A, Mousannif H, Moatassime H A, et al. Data quality in internet of things[J]. *Journal of Network and Computer Applications*, 2016, 73(73): 57-81
[5] Qi Chen. Research on the feature analysis of electric power big data based on Hadoop[D]. Beijing, North China Electric Power University 2016
[6] Baoyu Li, Yu Dai. Status and challenges of big data processing technology in Smart Grid[J]. *Science and Technology Economic Market*,2018(04):12-13
[7] Qian Dang, Yu Qiu, Li Wei. Application of key technologies in smart grid data processing, *Information technology and Informatization*. 2020, 10(1), 217-219
[8] Vuran M C, Akan O B, Akyildiz I F. Spatio-temporal correlation: theory and applications for wireless sensor networks[J]. *Computer Networks*, 2004, 45(3):245-259

[9] Baydogan M G, Runger G. Time series representation and similarity based on local autopatterns[J]. *Data Mining and Knowledge Discovery*, 2016, 30(2):1-34

[10] Peng J, Xiaofeng X, Yan Q, et al. High-Dimensional Data Anomaly Detection for WSNs Based on Deep Belief Network[J]. *Chinese Journal of Sensors and Actuators*, 2019

[11] Nevat I, Divakaran D M, Nagarajan S G. Anomaly Detection and Attribution in Networks With Temporally Correlated Traffic[J]. IEEE/ACM Transactions on Networking, 2018, 26(1):131-144

[12] Laxhammar R, Falkman G, Sviestins E. Anomaly detection in sea traffic - A comparison of the Gaussian Mixture Model and the Kernel Density Estimator[C]. 2009 12th International Conference on Information Fusion, 2009, 756-763

[13] Fatourechi M, Ward R K, Mason S G, et al. Comparison of Evaluation Metrics in Classification Applications with Imbalanced Datasets[C]// Seventh International Conference on Machine Learning & Applications. IEEE Computer Society, 2008