Proceedings of the 39th Chinese Control Conference
July 27-29, 2020, Shenyang, China

Online Clustering based Fault Data Detection Method for Distributed PV Sites

Shujie Wang[1], Feng Gao[1], Jiang Wu[1], Chao Zheng[1], Xingbo Fu[1], Fangwei Duan[2]

1. Ministry of Education Key Lab for Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an 710049, China
E-mail: {wangsj,zhengchao5025125,fxb123}@stu.xjtu.edu.cn, fgao@mail.xjtu.edu.cn, jiangwu@xjtu.edu.cn

2. Electric Power Research Institute, State Grid Liaoning Electric Power Co.Ltd., Shenyang 110000, China
E-mail: dfw8906@163.com

Abstract: In distributed photovoltaic (PV) sites, fault data detection is critical to ensure the safety of power grid. Accurate and reliable PV data is the basis of PV power generation performance analysis and power load forecasting. However, many PV power sites have high proportion of fault power measured data, which greatly impairs the analysis of power site performance. This paper summarizes three typical fault data types of PV data based on engineering experience. Utilizing Spark Streaming and k-means algorithm, a new method, namely the streaming k-means method under different time windows is adopted to detect the fault PV data in real time. In the meanwhile, the specified Silhouette Coefficient is used to choose the proper clustering number in each detection period. And in order to better display the clustering results, principal components analysis (PCA) is applied to present the data distribution in real time. In the numerical simulation, the actual data from Wuxi Hongdou PV power cites and the artificially generated data set are utilized to verify the proposed method. The experiment results show that the streaming k-means method can effectively identify various types of fault data and has a better detection rate than the 3-sigma recognition method and logistic regression.

Key Words: Distributed PV system, Fault data detection, Clustering

1 Introduction

In recent years, with the improvement of environmental awareness, new energy, especially renewable energy such as distributed PV system, has received more and more attention. And the installed capacity of distributed PVs has exploded. As a result, PV data has reached the level of big data. However, PV power measurement data will be abnormal due to a variety of causes in the engineering environment such as equipment failure, artificial limiting electricity, communication failure, PV arrays failure and large fluctuations in weather factors. At the same time, if fault data is not diagnosed in time, it will seriously affect the quality of data applications such as power planning or power forecasting. Therefore, it is of great significance to propose an online monitoring method for fault data of distributed PV systems.

Fault data detection is an important step in data processing. There are some common methods like rough set theory [1], principal component analysis (PCA) [2], cluster analysis [3] and quartile method [4]. But these common methods of fault data detection are not suitable for PV data because of uncertainties in PV data.

At present, most projects use probabilistic statistics to detect fault data. For example, the author in [5] uses the 3-sigma principle to detect fault PV data under the assumption that the probability distribution function of PV power data is normally distributed. However, it has been proved that the probability distribution of PV power is close to the normal distribution only on sunny days [6]. The author in [5] proposed a combined model based on quartile method and cluster analysis. The model is lack of validity of the fault data detection because it only measures the effect of detecting fault data by the error value of the equivalent power curve. [7] proposed a probabilistic power curve based on the Copula function to describe the relationship between solar irradiance and PV power. Then, they proposed corresponding fault data detection models for different characteristics of measured PV data. However, this model does not work without real-time solar data.

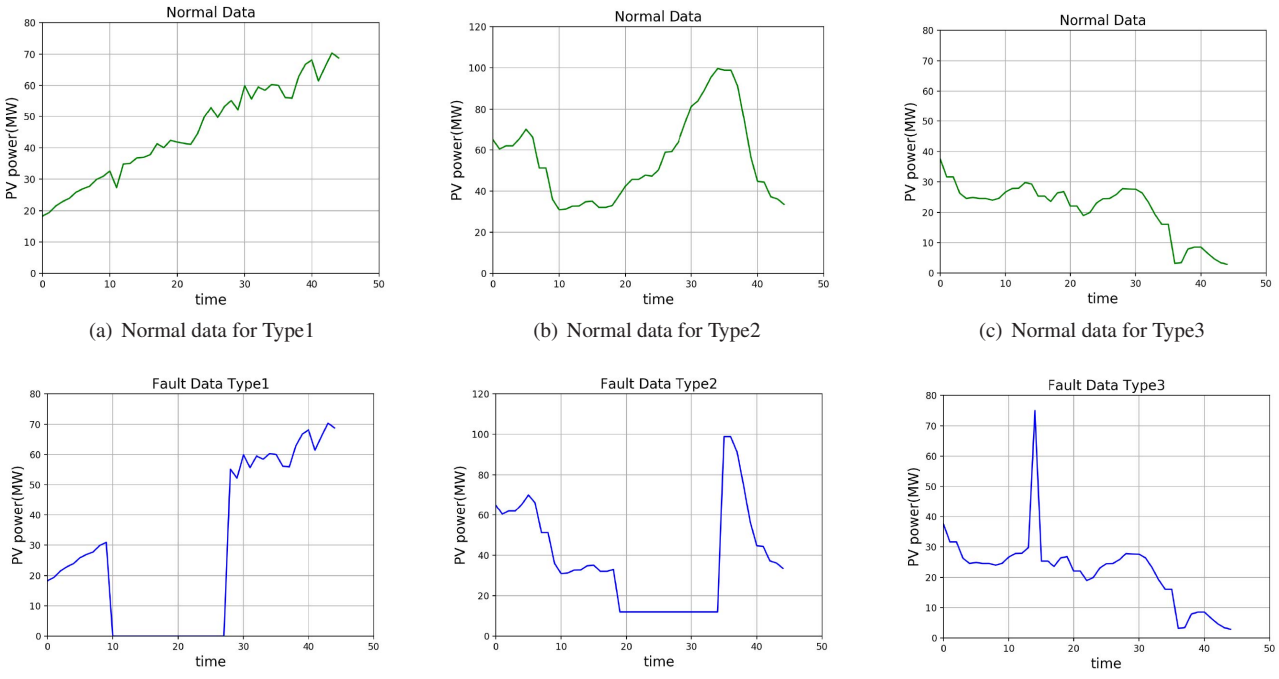
In this paper, fault PV data are classified three types. However, the three types of fault data cannot be directly detected by the conventional statistical data-based fault data detection methods (such as the 3-sigma fault data detection method) although they seem simple. The power of PV sites has a strong dependence on solar situations and is greatly affected by weather, which results in different power probability curves under different weather. Therefore, it is not guaranteed that the sample data in each period is distributed normally, which leads to a lower accuracy of identification using methods such as the 3-sigma principle. A new online k-means clustering method for distributed PV data is achieved based on Spark Streaming and k-means algorithm. The number of clusters is selected by Silhouette Coefficient (SI Coefficient) and PCA is used to visualize the clustering results in real time. Finaly a set of evaluation indicators for distributed PV fault data detection is proposed.

The rest of the paper is organized as follows. In Section 2, the classification of fault data, the Spark Streaming, the k-means algorithm and the PCA are introduced. Section 3 presents the method of choosing clustering number and the streaming k-means algorithm are given. And Section 4 provides relevant experiment results and analysis. Some conclusions are given in Section 5.

2 Preliminary

In this section, fault data of PV cites are summarized into three different types according to actual engineering experience. And the streaming k-means technology based on big

This work was supported by the State Grid Corporation of China under Science and Technology Project of "the research and development of the key technology of production simulation for the future high proportion new energy consumption adapting to the multi-region and multi-point distribution of energy in China".



(a) Normal data for Type1 (b) Normal data for Type2 (c) Normal data for Type3

(d) Fault Type1 (e) Fault Type2 (f) Fault Type3

Fig. 1: Normal Data and Fault Data

data processing and PCA technology for dimension reduction visualization are introduced.

2.1 The Classification of PV Fault Data

Based on previous research on [1],[4], this paper summarizes the three types of PV fault data and their causes:

- Type 1: The main characteristic of this type is that the measured PV power data remains at 0 or close to 0 in a continuous period of time when the solar irradiance is much greater than 0 as shown in Fig.1.(a) and Fig.1.(d); communication failure, measurement equipment failure or PV arrays failure will cause the failure.
- Type 2: The main characteristic of this type is that the measured data remains at a value which is lower than the right value but not 0 or close to 0 in a continuous period of time as shown in Fig.1.(b) and Fig.1.(e), and does not change with the solar irradiance; the main reason is PV power limitation and communication or measuring equipment failure;
- Type 3: The main characteristic of this type is that at a certain time, the measured PV power value is much higher than the right power data as shown in Fig.1.(c) and Fig.1.(f), and the value is normal before and after the moment, which means an abnormal spike is formed.And the remaining time of this type is usually short; the main reason is that the sensors of the communication or measurement equipment have been interfered by some reasons.

2.2 Spark Streaming and k-means

Due to the large amount of data and real-time requirement of distributed PV fault data online monitoring, a tool that can process a large amount of streaming data is urgently needed. Spark Streaming [8, 9] is an extension of Spark's core API that can achieve real-time streaming data processing with fault tolerance and high-throughput, which meets the requirements of the study. Spark Streaming can receive real-time input data from various sources, such as Kafka, Flume and HDFS. The processing structure can be stored in HDFS, DataBase, and SQL databases after processing. In this paper, the input data source is Kafka [12] and the clustering results are stored in SQL [13].

Based on the real-time processing function of Spark Streaming, this paper uses k-means algorithm [10] to finish the clustering. The k-means algorithm can be summarized in two steps: the first one is an assignment step and the second one is a refinement step. In the assignment step, we firstly randomly select k cluster centers and then we compute the euclidean distance between each data point di and those cluster centers. After that, data points are grouped into its nearest cluster which can be represented by the nearest centroid. In the refinement step, the k centroids are updated using the following formula:

$$c_k\left(t+1\right)=\frac{\sum_{i=1}^{N}\mu_k^i\left(t+1\right)\times d_i}{\sum_{i=1}^{N}\mu_k^i\left(t+1\right)} \tag{1}$$

where c_k is the $k-th$ centroid.

$$\mu_k^i=\begin{cases}1, & if\ d_i\in Cluster\,k\\ 0, & if\ d_i\notin Cluster\,k\end{cases} \tag{2}$$

After the centroids converge or the number of iterations reaches the specified number of times, the algorithm ends.

2.3 PCA

The power data of the PV cite is measured at an interval of about five minutes, which means that the feature quantity of the PV data for a day will reach 288. There is a problem that it is easy to fall into a dimensional disaster. As the

time window for the clustering of PV power data increases, the dimension of the vector also become higher. In order to facilitate data visualization, this paper introduces PCA [11] technology.

The specific method is as follows:

(1) Calculate the data covariance matrix and obtain the eigenvalues of the covariance matrix.
(2) Sort the eigenvalues according to size. The largest eigenvalue is the first principal component, the second largest eigenvalue is the second principal component, and so on.
(3) Calculate the feature vector, convert the feature vector into a unit feature vector and use the feature vector corresponding to the principal component as the transformation matrix. Use the data matrix to multiply the transformation matrix to achieve the principal component mapping.

3 Online Monitoring for PV Fault Data

In this section, the paper uses big data processing named Spark Streaming to do online k-means clustering under dynamic time window on power data of distributed PV cites based on the three types of PV anomaly data proposed in 2.1. Then, the number of clusters based on the SI Coefficients is selected. And a set of evaluation indicators suitable for distributed PV fault data detection is proposed.

3.1 Deciding n-cluster k utilizing SI Coefficient

Clustering is unsupervised learning and the number of categories needs to be set in advance. However, it is impossible to know whether there are fault data in the real-time data or how many clusters of fault data there are in the process of clustering real-time data. So the category parameter k cannot directly determined in the streaming k-means algorithm. This paper hopes to determine the true number of clusters from the data itself. Therefore, the clustering quality evaluation index named SI Coefficient is introduced as a factor for determining the number of clustering categories k under each time window.

The SI Coefficient of a sample point X_i is defined as follows:

$$S = (b - a)/max(a, b) \tag{3}$$

where a is the average distance between X_i and other samples in the same cluster, called the degree of aggregation, and b is the average distance between X_i and all samples in the nearest cluster, called the degree of separation. And the nearest cluster is defined as:

$$C_j = \arg\min_{C_k} \frac{1}{n} \sum_{p \in C_k} |p - X_i|^2 \tag{4}$$

where p is a sample in a certain cluster C_k. That is, the average distance from X_i to all samples in a cluster is used as a measurement of the distance from the point X_i to the cluster, and the cluster closest to X_i is selected as the nearest cluster.

Then, we average all the SI Coefficients after obtaining the SI Coefficient of all sample points. The range of the average SI Coefficient is [-1,1]. When the distance between samples within a class is closer and the distance between samples within different classes is farther, the average SI Coefficient and the clustering quality will be greater.

Therefore, k with the largest average SI Coefficient is the optimal number of clusters. In this paper, the selection value of k is 2,3,4. The maximum number of selectable categories is set to 4 in order to avoid three types of fault data appearing at the same time. Then, the minimum number of selectable categories is set to 2 because clustering is meaningless when the k is 1.

But this method does not work well when $k = 1$. If all the PV power data is normal in this time window, it will not be a good clustering result that k is not 1. So in this paper, some experiments are performed to decide a proper threshold of SI Coefficient τ to exclude the bad results before the streaming k-means algorithm.

$$\begin{cases} k = 1, & if\ SIcoefficient < \tau \\ k \in \{2, 3, 4\}, & if\ SIcoefficient \geq \tau \end{cases} \tag{5}$$

The optimal k will be selected to get the best clustering effect in each time window.

3.2 The Streaming k-means Technology under Time Window

In fact, the weather conditions of each PV cite are approximately the same for distributed PV cites in a factory under the same geographical conditions. Therefore, ideally, the power data of each PV cite has strong consistency. The method of online clustering of power data of multiple PV cites can be used to ignore the impact of weather changes on data and make it easier to detect fault data. In the meanwhile, for the goal of processing data in real time, Spark Streaming can be considered.

Spark Streaming combines the traditional k-means algorithm to achieve clustering of streaming data, namely the streaming k-means algorithm. The streaming k-means is an extension based on streaming data. In the streaming environment, data comes in batches and each batch contains the latest data points. The streaming k-means algorithm executes the k-means algorithm for each batch of new data. It assigns all new data points to the cluster closest to it and then updates the historical clustering center point.

The formula for updating the cluster center is:

$$c_{t+1} = (c_t n_t \alpha + x_t m_t)/(n_t \alpha + m_t) \tag{6}$$

$$n_{t+1} = n_t + m_t \tag{7}$$

where c_t is the cluster center after the last clustering. n_t is the number of all points accumulated. x_t is the new cluster center after adding the data of the current batch. m_t is the number of points in the current batch. The decay factor α is applied to the current point as a weighted discount when clustering new data. α ranges from 0 to 1. The larger α is, the greater the impact of historical data on new data is. When α is 1, each batch of data is given the same weight. This means that historical and new data have the same impact. When α is 0, the cluster center is completely determined by the new data and the historical data is ignored. In this experiment, α is set to 0. That is, the k-means algorithm is executed only for each batch of newly arrived data points and history data is ignored.

Then the formula is simplified to:

$$c_{t+1} = x_t \tag{8}$$

It means that the updated cluster center is completely obtained from the latest batch of data through the streaming kmeans algorithm. The clustering results of each the streaming k-means algorithm are also generated only by the latest batch of data.

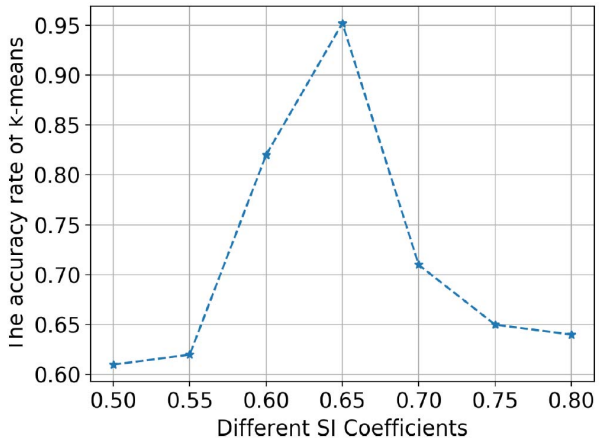


Fig. 2: The accuracy rate under different SI Coefficients

Spark Streaming provides a set of window operations to perform statistical analysis on incremental updates of largescale data through sliding window technology. Window operations requires setting two parameters:

- Window length: The length of the time window.
- Slide interval: The length by which the time window moves forward.

The time window size of the online monitoring data is set by setting the window length variable. In theory, different time window lengths may contain different time series information. The shorter the time window is, the greater the difference in data distribution is and the more sensitive it is to fault data Type 3. The longer the time window is, the easier it is to catch fault data Types 1 and 2 over duration. In this paper, the window length and the slide interval are both set as one hour according to historical experience. The streaming k-means parameter k is set to 2, 3, and 4 to do online clustering of the power data at the set window length and slide interval. Then, the SI Coefficient is used to select the optimal k and the clustering result is used as a valid result of the fault data detecting system.

4 Case Analysis

A method of data normalization is introduced to keep data of all sites in the same range. At the same time, an experiment to choose proper threshold for SI Coefficient to increase the clustering quality is performed. At last, the comparative experiments under 3-sigma, logistic regression and the streaming k-means algorithm are performed to verify the effectiveness of proposed the streaming k-means algorithm.

4.1 Experimental Enviroment and Data Description

The big data distributed environment of this method includes three nodes, a master and two slaves. The operating system of the experimental environment is centos 6.0, and the software configuration includes jdk, hadoop, spark, kafka, zookeeper and scala.

The data set is the actual PV power data collected at 30 PV cites in a PV industrial garden in Wuxi. The distribution of these PV cites is tight and the weather conditions are basically the same, which meets the experimental requirements. The time granularity of the data is five minutes and there are 288 points in a day. The distribution of fault data in this data set is relatively scattered. So, this paper uses the normal original data without errors and gets the test data set by artificially adding fault data in the normal data to facilitate the experiment. In order to test the performance of the clustering method, this paper selects four days of PV data to test the fault data detection algorithm proposed.

4.2 Data Normalization

The installed capacity of each PV cite is different. So, the data must be standardized firstly to eliminate the impact of the installed capacity on the detecting results. The ordinary normalization method will weaken the characteristics of the fault data and reduce its difference from the normal data. Then, increase the difficulty of detecting the fault data. Therefore, this paper normalizes the data by the following formula:

$$X_i = x_i * \frac{c_i}{c_m} \quad i = 0, 1, 2...n \tag{9}$$

where:

- X_i is the normalized data of the i-th PV cite.
- x_i is the original data of the i-th PV cite.
- c_i is the installed capacity of the i-th PV cite.
- c_m is the median value of the installed capacity of all sites.

4.3 Choosing Proper Threshold for SI Coefficient

As mentioned in section 3.1, a proper threshold for SI Coefficient needs to be choose before deciding the number of cluster. The threshold will help to determine the number of clustering for increaseing the clustering quality. This experiment sets that the value of SI Coefficient threshold ranges from 0.5 to 0.8. Then, we compare the accuracy rate of streaming kmeans algorithm to select the best threshold.

According to the Fig.2, the threshold of SI Coefficient will be set as 0.65. If the corresponding SI Coefficient is below the threshold, the k will be set as 1. In other situations, the k will be the value which obtains the highest SI Coefficient.

4.4 Experimental Results and Analysis

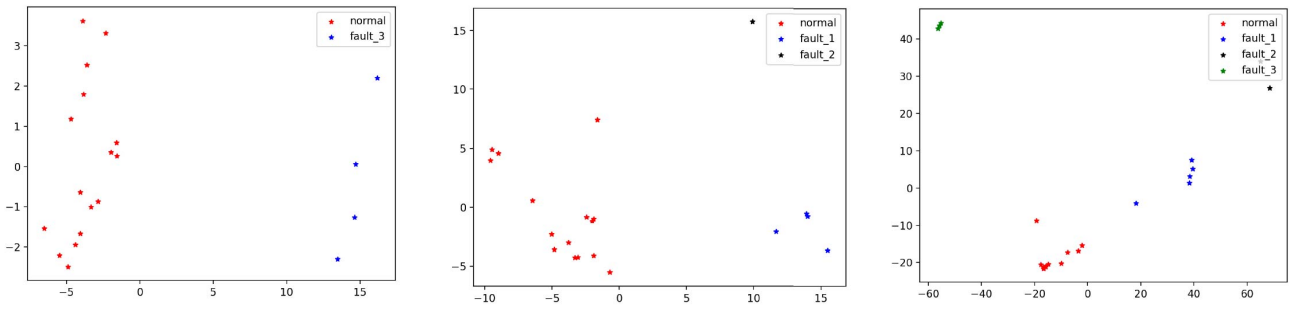
The clustering visualization results utilizing PCA are shown in this section. And the effectiveness of the streaming k-means algorithm is also verfied by comparing with 3sigma method and logistic regression method.

Clustering experiments are performed on the different situations that one, two or three types of fault data occurs at the same time. The Fig.3 are the visualization results with PCA dimension reduction.

In Fig.3, there are two, three or four types of PV data, the clustering results show that all power sites can all be clustered correctly.

In this paper, 3-sigma principle and logistic regression are used as comparative experiments.

The 3-sigma method is used to detect on the data after



(a) k=2 (b) k=3 (c) k=4

Fig. 3: Clustering results:(a)k=2(b)k=3(c)k=4

Table 1: The detection results under 3-sigma, the streaming k-means and LR

the method	2018/7/26		2018/8/15		2018/8/19		2018/8/25		b/%
	r/%	b/%	r/%	b/%	r/%	b/%	r/%	b/%	
3-sigma	64.32	3.21	76.92	3.96	17.33	4.01	100	3.33	3.62
logistic regression	92.45	6.15	84.72	36.51	41.73	10.8	53.91	13.67	16.78
the streaming k-means	92.86	2.46	100	2.78	84.62	1.85	100	2	2.27

normalization at each moment of each cite. The 3-sigma principle can be expressed as:

$$X_i = \begin{cases} normal, & if\ X_i \in (\overline{X} - 3 * \sigma) \\ abnormal, & if\ X_i \notin (\overline{X} - 3 * \sigma) \end{cases} \quad (10)$$

where, $i = 1, 2, 3...n$, X_i is the $i-th$ PV cite data, n is the number of PV cites, σ is the standard deviation of PV cites data, $\overline{X}$ is the mean value of PV cites data.

And logistic regression is a classification algorithm used to assign observations to a discrete set of classes. Logistic regression transforms its output using the logistic sigmoid function to return a probability value which can then be mapped to two or more discrete classes.

To compare the detection effectiveness of the two methods above and our proposed streaming k-means algorithm, this paper chooses four-day PV data in Wuxi Hongdou PV power plants, which contain different weather conditions. The detection results are shown in Table.1.

Among them, the accuracy rate r refers to the proportion of the recognized fault data of three categories in the test fault dataset.

$$r = \frac{n_{re}}{n_{te}} * 100\% \quad (11)$$

where, n_{re} is the number of fault data correctly identified by the model and n_{te} is the actual number of fault data.

To a certain extent, the high accuracy can demonstrate the effectiveness of the recognition model. But the error rate of the model is still critical.

$$\zeta = \frac{n_{we}}{n_t} * 100\% \quad (12)$$

where, the mis-identification rate ζ refers to the proportion of the total data in a dataset that is misidentified by other data in three categories. n_{we} is the number of data incorrectly identified and n_t is the number of total data.

According to the results in Table.1, the streaming kmeans algorithm gets a higher accuracy rate and a lower misidentification rate than the other two methods. It is also demonstrated that the streaming k-means algorithm can not only process PV data in real-time but can reduce misidentification rate to increase the quality of data.

5 Concolusion

The installed capacity of PVs has exploded with the development of distributed PVs. The probability of fault PV data is rising. So, fault data detection for PV power data has become a crucial issue. Due to its own characteristics, PV power data has a strong uncertainty, which is difficult to detect using traditional fault data detection methods. And most of the current researches are off-line detection method, ignoring the high real-time requirement of fault data detection.

In this paper, an online monitoring method for fault data detection of distributed PV systems is proposed. PV fault data is divided into three types and online k-means clustering of distributed PV data is performed based on Spark Streaming and the number of clusters is selected by SI Coefficient. Then, this paper proposes a set of evaluation indicators suitable for distributed PV fault data detection to evaluate the clustering results. Experimental results show that the proposed method has higher accuracy rate and lower misidentification rate than traditional fault data detection methods. And this method achieves online detection unlike the traditional offline detection method. It can detect false PV data timely and effectively, guarantee data quality and grid security, and improve the stability of the power system.

References

[1] X. Zhang, Vibration fault diagnosis of hydropower units based on rough sets and multi-class support vector machines, *Proceedings of the CSEE*, 30(20): 88-93, 2010.

[2] Saha, Snake validation: a PCA-based outlier detection method, *IEEE Signal Processing Letters*, 16(6): 549-552, 2009.

[3] P. Yang, An outlier detection algorithm based on spectral clustering, *IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application*, 2008: 507–510.

[4] Y. Zhao, Characteristics and treatment methods of abnormal

wind data clusters in wind farms, *Automation of Electric Power Systems*, 38(21): 39-46, 2014.

[5] Y. Zhao, Outlier detection rules for fault detection in solar PV arrays, *in Proceedings of 5th International Conference on Control and Automation*, 2013: 2913-2920.

[6] W. Zhao, Probability distribution error estimation method of conditional prediction error for PV power generation, *Automation of Electric Power Systems*, 39(16): 8-15, 2015.

[7] Y. Gong, Machine identification algorithm based on Copula theory for PV power high ratio anomaly data, *Automation of Electric Power Systems*, 40(9): 16-22, 2016.

[8] A. Bifet, Streamdm: Advanced data mining in spark streaming, *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, 2015: 1608-1611.

[9] L. Rettig, Online anomaly detection over big data streams, *Applied Data Science*, 2019: 289-312.

[10] K. Krishna, enetic K-means algorithm, *IEEE Transactions on Systems, Man, and Cybernetics*, Part B (Cybernetics) 29.3:433-439, 1999.

[11] S. Wold, Principal component analysis, *Chemometrics and intelligent laboratory systems*, 1987: 37-52.

[12] L. Noac, A performance evaluation of Apache Kafka in support of big data streaming applications, *2017 IEEE International Conference on Big Data (Big Data)*, 2017: 4803-4806.

[13] M. Armbrust, Spark sql: Relational data processing in spark, *Proceedings of the 2015 ACM SIGMOD international conference on management of data*, 2015: 1383-1394.