Bad Data Detection Algorithm for PMU Based on Spectral Clustering

Zhiwei Yang, Hao Liu, Tianshu Bi, *Senior Member*, *IEEE*, and Qixun Yang

***Abstract*—Phasor measurement units (PMUs) can provide real-time measurement data to construct the ubiquitous electric of the Internet of Things. However, due to complex factors on site, PMU data can be easily compromised by interference or synchronization jitter. It will lead to various levels of PMU data quality issues, which can directly affect the PMU-based application and even threaten the safety of power systems. In order to improve the PMU data quality, a data-driven PMU bad data detection algorithm based on spectral clustering using single PMU data is proposed in this paper. The proposed algorithm does not require the system topology and parameters. Firstly, a data identification method based on a decision tree is proposed to distinguish event data and bad data by using the slope feature of each data. Then, a bad data detection method based on spectral clustering is developed. By analyzing the weighted relationships among all the data, this method can detect the bad data with a small deviation. Simulations and results of field recording data test illustrate that this data-driven method can achieve bad data identification and detection effectively. This technique can improve PMU data quality to guarantee its applications in the power systems.**

***Index Terms*—Phasor measurement units (PMUs), bad data detection, event data identification, decision tree, spectral clustering.**

I. Introduction

PHASOR measurement units (PMUs) have become an important mechanism used in the ubiquitous electric of the Internet of Things to achieve state perception, due to their rapidity, synchronism, and accuracy [1]. Furthermore, PMUs can provide real-time phasor time data for critical power system applications such as remedial action schemes, oscillation detection, and state estimation [2] - [6]. Up to 2018, approximately 3000 PMUs have been installed and put into operation in China, covering the majority of 220 kV and above substations, power plants, and grid-connected renewable energy collections [7]. In addition, according to the statistics in 2017, it is reported that around 2500 commercial PMUs have been installed in North America [8].

However, in view of the complex factors, PMU data is vulnerable to many factors [9]. For example, a jitter of globle positioning system (GPS) signal can cause phase angle deviation. It is also possible that PMU data may have a spike due to an interferent, or a mistake of data transmission. Such issues lead to various degrees of data quality issues in PMU data. According to the 2011 Five-Year Plan issued by the California Independent System Operator (ISO), around 10% to 17% of PMU data in North America experience problems [10]. As discussed in [11], around 20%-30% of PMU data in China experience data quality problems. Data quality issues make the system less observable, affect the performance of state estimation and parameter identification based on PMUs, and even threaten the safe and stable operation of power systems. The detection of PMU bad data has become a critical issue and plays an important role in improving data quality and ensuring accurate state perception.

Various methods have been proposed to detect bad data in the power systems. In [12], a new approach for identifying measurement errors in DC power flow is presented by exploiting the singularity of the impedance matrix and the sparsity of the error vector. It leverages the structure of the power system and can compute the measurement errors accurately. In [13], a bad data detection method is presented based on state estimation. The phasor-measurement-based state estimator improves data consistency by identifying angle biases and current scaling errors. A time-series prediction model combined with Kalman filter and smoothing algorithm to clean the bad data is introduced in [14]. Reference [15] proposes a method based on the unscented Kalman filter in conjunction with a state estimation algorithm to detect bad data in real-time. According to [16], bad data from faulty current transformers can be detected by a linear weighted least square-based state estimation algorithm. Reference [17] proposes a robust generalized estimator to detect bad data by exploiting the temporal correlation and the statistical consistency of measurements. Both state estimator and Kalman filter method require system topology and line parameters with multiple PMU measurements. Therefore, the results of both methods will be affected in cases where an error exists in the topology or parameter of the system.

Some data-driven methods have been proposed to detect data anomaly. Traditional methods for bad data detection are based on the format of the sent data in the protocol. In [18],

Manuscript received: April 12, 2019; accepted: January 13, 2020. Date of CrossCheck: January 13, 2020. Date of online publication: April 9, 2020.

This work was supported by the National Key R&D Program (No. 2017YFB0902901) and the National Natural Science Foundation of China (No. 51627811, No. 51725702, and No. 51707064).

Z. Yang, H. Liu (corresponding author), T. Bi, and Q. Yang are with the State Key Laboratory of Alternate Electrical Power System with Renewable Energy Sources, North China Electric Power University, Beijing 102206, China (e-mail: hsyangzhiwei@163.com; hliu@ncepu.edu.cn; tsbi@ncepu.edu.cn; yangqixun@sf-auto.com).

DOI: 10.35833/MPCE.2019.000457

a selection of detection criteria based on logical judgments is developed. If the data exceeds the set threshold, it is considered to be bad data. However, if there is a large disturbance in the power system, the specific threshold set in advance makes no sense. Measurement information in multiple PMUs is used in [19], where an online data-driven approach is introduced for the detection of low-quality phasor measurements based on spatiotemporal similarities among multiple-time-instant synchrophasor measurements. Similarly, the low-rank property of the matrix and the sparsity of the anomalies are used to detect bad data in [20]. In [21], a method based on principal component analysis is proposed to separate signals into low-dimensional feature components and high-dimensional noise components for bad data detection. These methods utilize the information of multiple PMU measurements to achieve bad data detection.

In some areas, only a small number of PMUs are present and the information of multiple PMU measurements is difficult to obtain and single PMU measurement can only be used to achieve the detection. Reference [22] develops an ensemble learning algorithm based on a single PMU with three simple models to detect anomaly data. An alternative density-based clustering method is proposed to cluster the phasor data to detect bad data for classification in [23]. Reference [24] presents machine learning techniques based on the support vector machine for bad detection. These methods are based on a single PMU, and when the data such as step data is presented during the events, the methods may not be suitable.

In this paper, a data-driven PMU bad data detection algorithm by a single PMU measurement is proposed which is based on spectral clustering. In order to distinguish event data from bad data, a bad data and event data identification method based on a decision tree is first developed which utilizes the slope feature of each data. Then a subsequent bad data detection method based on spectral clustering is proposed, which can detect bad data with small deviation values by the weight among the data. The proposed algorithm does not require the system topology of parameters. It can avoid the misjudgment of event data. The feasibility and accuracy of the proposed method are verified through simulations and field recorded data. The results show that this data-driven method can achieve bad data identification and detection effectively. It can guarantee better application of PMU data.

The remaining sections of this paper are organized as follows. In Section II, a bad data and event data identification method based on the decision tree is proposed. Section III details a detection method for bad data based on spectral clustering. The results of the numerical experiments on simulation and field PMU data are documented in Section IV. Finally, Section V concludes the paper.

II. Identification Method of Bad Data and Event Data

A. Features of Bad Data and Event Data

This paper mainly studies the PMU bad data which is affected by interference or jitter. These bad data deviate from the normal values. By analyzing a large amount of field data, most of the bad data exists alone and the number of contiguous bad data is no more than three. It is also pointed that the outliers are all isolated and not in sequence in [22], [25]. Meanwhile, the amplitude is taken as an example to introduce this method. It can be applied to amplitude, frequency, and rate of change of frequency, where the amplitude includes voltage amplitude and current amplitude. But it is not suitable for the phase angle, because when the frequency is offset, the phase angle changes from −180° to 180°.

The schematic in Fig. 1 includes some bad data in the steady state. The grey circle represents normal data. The blue circles show anomaly data with higher amplitude and the red circles show anomaly data with a smaller amplitude. The number of contiguous bad data may be one, two, or three as shown in Fig. 1(a), (b), and (c), respectively. Taking Fig. 1(a) as an example, two possibilities of bad data can be seen in which the amplitude may be larger or smaller than normal value. Similarly, the possibilities in Fig. 1(b) and Fig. 1(c) are four and eight, respectively.

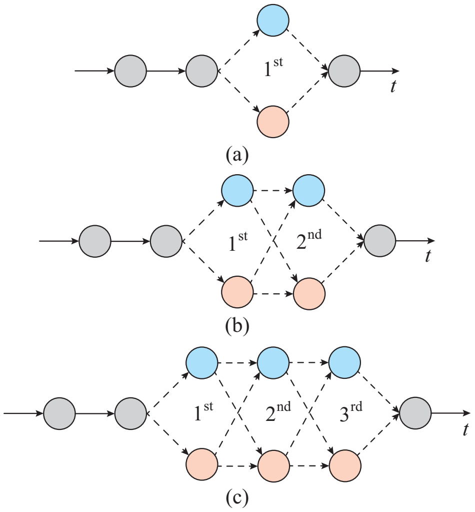


Fig. 1. Schematic of possible bad data. (a) One bad data. (b) Two bad data. (c) Three bad data.

For the purpose of avoiding misjudgment of event data, a comparison is carried out between event data, bad data, and normal data, as illustrated in Fig. 2. In Fig. 2, $|X_i|$ is the amplitude of each data. As shown in Fig. 2(a), when $t=t_{i+1}$, the amplitude step occurs and the yellow circle represents the step data. The data X_i is defined as a step point, in which the data before X_i and the data after X_{i+1} can be both considered as normal data. Figure 2(b) represents the possibility of three contiguous bad data events. The blue circle represents bad data whose amplitudes are higher than the normal value and close to each other. Additionally, Fig. 2(c) represents the normal data. According to the comparison, the difference between event data and bad data is the number of contiguous data points with close amplitudes. In this case, the number of event data is more than three, and the number of contiguous bad data is three or less. Thus, the method is able to distinguish them based on the features of four contiguous data.

The slope of each data k_i is calculated by (1).

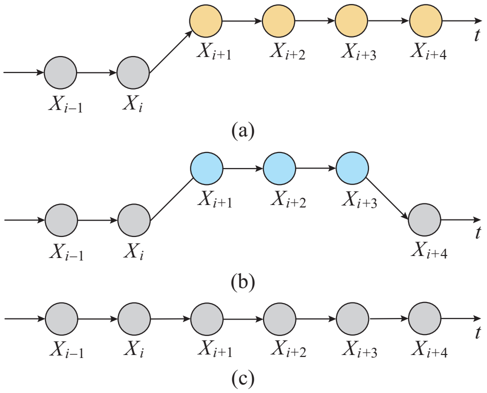


Fig. 2. Comparison between event data, bad data, and normal data. (a) Event data. (b) Bad data. (c) Normal data.

$$k_i=\left|\frac{|X_{i+1}|-|X_i|}{t_{i+1}-t_i}\right| \tag{1}$$

When the power system is in normal operation, the data is ambient with a small slope. However, when there is event data or bad data, the amplitude changes and the slopes become larger. A comparison of the slopes of event data and bad data in Fig. 2 is provided in Table I.

TABLE I
COMPARISON OF SLOPE VALUES OF EVENT DATA, BAD DATA, AND NORMAL DATA

Data type	k_i	k_{i+1}	k_{i+2}	k_{i+3}
Event data	Large	Small	Small	Small
Bad data	Large	Small	Small	Large
Normal data	Small	Small	Small	Small

Table I shows that the slopes are large, small, small, and small when starting from the step point X_i. The contiguous four-point slope of normal data or bad data does not change in this manner. Therefore, the slopes of four contiguous data from the step point have a unique feature that can be used to detect when the step occurs.

When the amplitude step occurs, it is difficult to calculate the value of the amplitude step without the parameter of lines. This means it is difficult to determine the threshold of the slope value of the step point. It is hard to detect the step point by using the threshold judgment method. Therefore, this paper proposes an identification method based on decision tree first, which avoids the subjectivity of artificially setting thresholds through the training of a large amount of field data. On this basis, the bad data is further detected. When the system oscillates, the identification method is still applicable, and this process is verified in Section IV.

B. Construction of Decision Tree

The identification of the event data and non-event data can be equivalent to a binary classification issue. A machine learning method based on the C4.5 decision tree is an effective tool to solve this problem [26]. This tool works well as it uses the information gain ratio to select features rather than the information gain in the ID3 algorithm, avoiding the preference for features with more values. Also, there are many other similar machine learning algorithms like the random forest, pre-pruning decision tree and classification, and regression tree, etc. The random forest consists of multiple decision trees. It has obvious advantages when dealing with large sample or high-dimensional feature data. In this paper, the sample data in the identification problem is small. There is no need to use the random forest algorithm. The pre-pruning decision tree can reduce the training time and test time. However, the branches of the tree constructed by C4.5 are only 4. Therefore, it does not need pre-pruning which might cause under-fitting. The classification and regression tree (CART) selects the best features by the Gini index, which is better for large sample data. The CART method is not necessary. Furthermore, a large number of simulations and field tests have proved that the C4.5 method has enough high accuracy, which can be seen in Section IV.

As shown in Fig. 2(a), the label of step point is $l=1$, while the others are $l=0$. The features of each data point are the slope values of the contiguous four data points including itself such as (k_i, k_{i+1}, k_{i+2}, k_{i+3}). Thus, there are a total of four features of each data point, recorded as $(a=k_i, b=k_{i+1}, c=k_{i+2}, d=k_{i+3})$. The construction of decision tree is then performed using a large amount of field data. In this method, 80% of all the data is randomly selected as the training set $\boldsymbol{D}$ including the event data and non-event data. 20% of the data is the test set $\boldsymbol{D}'$. The training data is used to construct a decision tree. The test data is used to verify its accuracy. The detailed steps are as followed.

The total information entropy of the training data $\boldsymbol{D}$ is calculated by:

$$Z(\boldsymbol{D})=-\sum_{i=1}^{2} z_i \log_2 z_i \tag{2}$$

where z_1 is the proportion of step point in $\boldsymbol{D}$; z_2 is the proportion of non-step point in $\boldsymbol{D}$; and $Z(\boldsymbol{D})$ is the uncertainty of the data label. The information entropy is one of the most commonly used indicators for measuring the purity of a sample.

Assume that the feature b is first selected to partition $\boldsymbol{D}$ and is discretized by dichotomy. Meanwhile, there are j different values in the feature b. Divide these values from small to large to form a collection $\{b^1, b^2, ..., b^j\}$. Set the median point of each interval $[b^i, b^{i+1})$ as the split point s_i. A split point collection S can be obtained by (3).

$$S=\left\{s_i=\frac{b^i+b^{i+1}}{2}\middle|1\le i\le j-1\right\} \tag{3}$$

Each split point can divide the training data $\boldsymbol{D}$ into subsets $\boldsymbol{D}_s^-$ and $\boldsymbol{D}_s^+$. $\boldsymbol{D}_s^-$ represents the collection of training data where $b_i\le s_i$, and $\boldsymbol{D}_s^+$ represents the collection of the training data where $b_i>s_i$. The information gain of s_i is calculated:

$$O(\boldsymbol{D},b,s_i)=Z(\boldsymbol{D})-\frac{|\boldsymbol{D}_{s_i}^-|}{|\boldsymbol{D}|}Z(\boldsymbol{D}_{s_i}^-)-\frac{|\boldsymbol{D}_{s_i}^+|}{|\boldsymbol{D}|}Z(\boldsymbol{D}_{s_i}^+) \tag{4}$$

where $|\boldsymbol{D}|$ is the number of data; $|\boldsymbol{D}_{s_i}^-|/|\boldsymbol{D}|$ is the weight of the data whose feature $b_i\le s_i$; and $|\boldsymbol{D}_{s_i}^+|/|\boldsymbol{D}|$ is the weight of the data whose feature $b_i > s_i$. The larger the information gain O is, the better effect the split point s_i has. The ID3 algorithm selects the maximum information gain, which has a preference for the features with more values. The C4.5 decision tree defines the gain ratio to select the optimal feature. The definition is as follows:

$$o(\boldsymbol{D},b,s_i)=\frac{O(\boldsymbol{D},b,s_i)}{I(b)} \tag{5}$$

$$I(b)=-\sum_{\beta\in\{-,+\}}\frac{\left|\boldsymbol{D}_{s_i}^{\beta}\right|}{|\boldsymbol{D}|}\log_2\frac{\left|\boldsymbol{D}_{s_i}^{\beta}\right|}{|\boldsymbol{D}|} \tag{6}$$

where $I(b)$ is the intrinsic value.

Select the maximum of gain ratio $o(\boldsymbol{D},b,s_i)$ as the gain ratio of the feature b. Therefore, select the split point s_b with the largest gain ratio $o(\boldsymbol{D},b,s_b)$ as the first branch node of the decision tree. The structure of the decision tree is shown as follows.

Figure 3 shows the detailed process of the decision tree. Set the decision tree depth p and the threshold of the information gain ratio ε, which determines the identification accuracy. The depth p represents the times of the recursive calculation. There are three situations where the decision tree ends. The first is that if the maximum time of calculations reaches p, the division is stopped. The second is that if all the information gain ratios of each feature are less than ε, it is not divided either. The third is that if all labels of one leaf nodes are the same, there is no more division. First, all the data X_i is input. The gain ratios of features (a, b, c, d) are calculated separately. Then, select the largest gain ratio to compare with ε. If the gain ratio is greater than ε, the corresponding feature is used as the feature of the first division. Suppose b as the selected feature. The split point s_b is called the branch node. The data X_i whose feature $b_i \le s_b$ is in one collection and that whose feature $b_i > s_b$ is in another collection. If the gain ratio is less than ε, the label of the data is the same and the tree is a single node tree. Repeat the above steps recursively until the labels of the data in one collection are the same. The last layer node is called the leaf node.

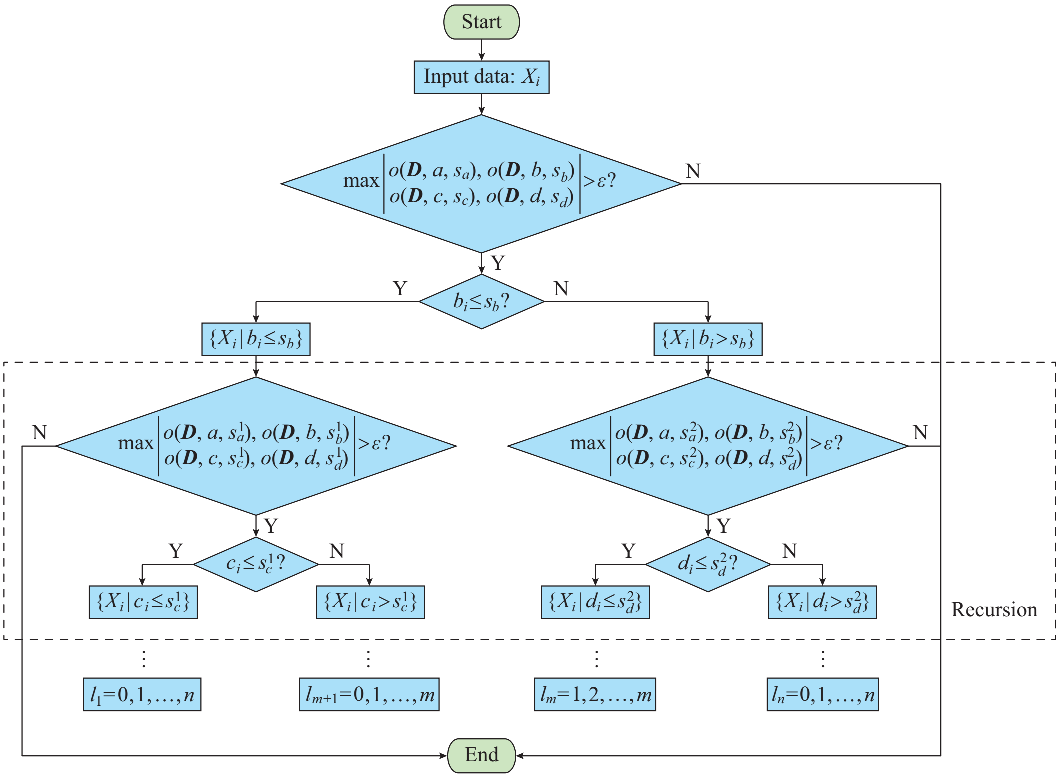

Start
Input data: X_i
$\max\begin{vmatrix} o(\boldsymbol{D},a,s_a), o(\boldsymbol{D},b,s_b) \\ o(\boldsymbol{D},c,s_c), o(\boldsymbol{D},d,s_d) \end{vmatrix} > \varepsilon$?
N
Y
Y
$b_i \le s_b$?
N
$\{X_i \mid b_i \le s_b\}$
$\{X_i \mid b_i > s_b\}$
N
$\max\begin{vmatrix} o(\boldsymbol{D},a,s_a^1), o(\boldsymbol{D},b,s_b^1) \\ o(\boldsymbol{D},c,s_c^1), o(\boldsymbol{D},d,s_d^1) \end{vmatrix} > \varepsilon$?
$\max\begin{vmatrix} o(\boldsymbol{D},a,s_a^2), o(\boldsymbol{D},b,s_b^2) \\ o(\boldsymbol{D},c,s_c^2), o(\boldsymbol{D},d,s_d^2) \end{vmatrix} > \varepsilon$?
N
Y
Y
Y
$c_i \le s_c^1$?
N
Y
$d_i \le s_d^2$?
N
$\{X_i \mid c_i \le s_c^1\}$
$\{X_i \mid c_i > s_c^1\}$
$\{X_i \mid d_i \le s_d^2\}$
$\{X_i \mid d_i > s_d^2\}$
Recursion
$l_1 = 0, 1, \ldots, n$
$l_{m+1} = 0, 1, \ldots, m$
$l_m = 1, 2, \ldots, m$
$l_n = 0, 1, \ldots, n$
End

Fig. 3. Flow chart of a decision tree.

A decision function is used to indicate whether the test data $\boldsymbol{D}'$ contains the step point. The test data is put into the decision tree which is suitably trained to judge its corresponding label l_i. The decision function is described as follows:

$$f(X_i')=\begin{cases}0 \\ 1\end{cases} \quad i=1,2,\ldots,k \tag{7}$$

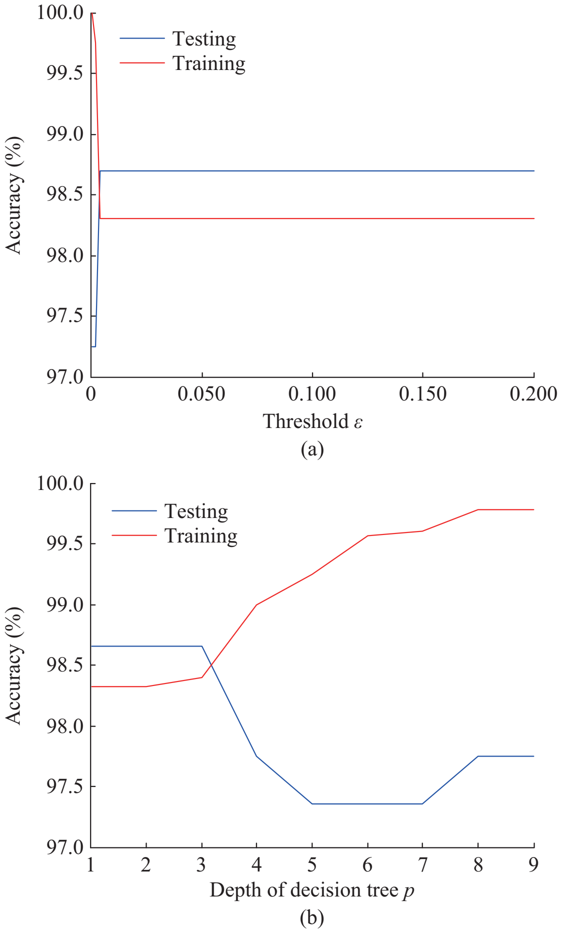
where X_i' represents the data in $\boldsymbol{D}'$. Equation (7) indicates that if there is any step point in $\boldsymbol{D}'$, the corresponding label should be 1 through the decision tree. The remaining non-step data labels are 0. It is like the step data X_m in Fig. 3.

Following this, the data before and after the step point are tested for bad data, so as to achieve the purpose of correctly distinguishing between bad data and event data.

C. Parameter Setting

In order to get better results, a threshold ε of the information gain ratio and the depth of the decision tree p should be set at the beginning. The optimal parameters as follows are obtained by traversing.

Figure 4(a) shows that when ε gradually increases to 0.0038, the accuracy of identification result obtained from the test data is up to 98.7% and then remains constant. Therefore, the threshold ε should be set to 0.0038. Figure 4(b) demonstrates that if the depth is greater than 3, the accuracy of the test data will decrease. The greater the depth of the decision tree is, the more complex the decision tree will be, which results in overfitting and reducing the test data accuracy. Thus, the depth of the decision tree p should be set as 3.



(a)

(b)

Fig. 4. Results of optimal parameters. (a) Relationship between ε and accuracy. (b) Relationship between p and accuracy.

III. Bad Data Detection Method

Event data can be successfully distinguished using the detailed process above. As PMU field data obeys a Gaussian distribution, the amplitudes of the data before the step occurs are shown in Fig. 2(a) with grey circles. The amplitudes of the data after the step occurs are filtered separately by the 3σ rule [27], as shown in Fig. 2(a) with yellow circles.

$$P\left(\left|\left|X_i\right|-\mu\right|\le 3\sigma\right)\le 99.73\% \tag{8}$$

where μ is the mean value of the amplitudes; and σ is the standard deviation of the amplitudes. If there is bad data, the bad data might be outside the range $(\mu-3\sigma,\mu+3\sigma)$ in Fig. 5.

In Fig. 5, the data distributed between $(\mu-\sigma,\mu+\sigma)$ is considered as normal data. The data out of $\mu-3\sigma$ and $\mu+3\sigma$ is confirmed as bad data. But for the data between $(\mu-3\sigma,\mu-\sigma)$ and $(\mu+\sigma,\mu+3\sigma)$, they can be good data or bad data, which cannot be detected by the 3σ rule. When the amplitude of the bad data is close to the mean value of the data set, they cannot be detected by this rule. Thus, a detection method based on spectral clustering is still needed.

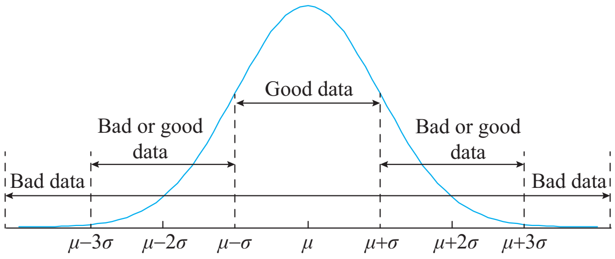


Fig. 5. Diagram of bad data distribution.

A. Spectral Clustering Theory

After the initial filtering of bad data by the 3σ rule, a spectral clustering method is developed to detect bad data. Unlike the density-based spatial clustering of applications with noise (DBSCAN) method in [23], a spectral clustering method is uniquely graph-based and transforms the clustering problem into a graph segmentation problem. For the purpose of minimizing the cost of segmentation, the undirected weighted graph composed of a single sample is divided into multiple subgraphs in order to implement the clustering of bad and normal data, as shown in Fig. 6.

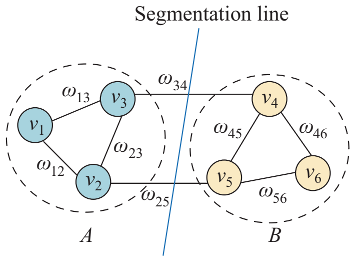


Fig. 6. Segmentation of graphs in spectral clustering.

Figure 6 shows a graph G in which the vertices v_i represent each data X_i in the sample. The blue vertices represent the normal data, the yellow vertices represent the bad data, and the edge represents the relationship between the two vertices. The relationship is the weight ω_{ij} which indicates the degree of similarity between v_i and v_j. As it is an undirected graph, $\omega_{ij}=\omega_{ji}$. The subgraph composed of normal data is called A, and that composed of bad data is B.

The purpose of spectral clustering is to cut the graph G to obtain two clusters: one with normal data, and the other with bad data. This requires the greatest similarity within the subgraph and the smallest similarity between sub-graphs, which is similar to the segmentation result of the blue line in Fig. 6. The total normal data is in subgraph A, and the total bad data is in subgraph B. The weights of the cut between A and B are defined as:

$$Cut(A,B)=\sum_{i\in A,j\in B}\omega_{ij} \tag{9}$$

Meanwhile, to maximize the number of vertices contained in each subgraph, the expression in (8) is improved as follows:

$$RCut(A,B)=\frac{1}{2}\left(\frac{Cut(A,B)}{|A|}+\frac{Cut(A,B)}{|B|}\right) \tag{10}$$

where $|A|$, $|B|$ are the numbers of vertices in subgraph A and |subgraph B, respectively. Extending the equation to m subgraphs, the objective function becomes:

$$RCut(A_1,A_2,\cdots,A_m)=\frac{1}{2}\sum_{i=1}^{m}\frac{Cut(A_i,\bar{A}_i)}{|A_i|} \tag{11}$$

Therefore, the objective function of spectral clustering is to solve the minimum value of (10), which is an NP-hard problem. It is transformed into the spectral decomposition problem of the similarity matrix. The suitable eigenvectors obtained by spectral decomposition are used to describe the low-dimensional structure of the data. The results are then obtained by using classical methods such as K-means.

First, the data in the sample is pre-processed, and the deviation r_i between the amplitude and the mean value is taken as the clustering feature of each data X_i by (12).

$$r_i=\left|\left|X_i\right|-\frac{1}{N}\sum_{i=1}^{n}\left|X_i\right|\right| \tag{12}$$

The similarity matrix $\boldsymbol{W}$ is established according to the similarity between any two data, and the similarity of any two data is defined as follows:

$$\omega_{ij}=\begin{cases}\exp\left(-\dfrac{\left\|r_i-r_j\right\|^2}{\delta^2}\right) & i\neq j\\ 0 & i=j\end{cases} \tag{13}$$

where δ is the scale parameter, which is set by the local scaling idea [28].

The degree matrix $\boldsymbol{D}_d$ is a diagonal matrix shown in (14).

$$\boldsymbol{D}_d=\begin{bmatrix} d_1 & 0 & \dots & 0\\ 0 & d_2 & \dots & 0\\ \vdots & \vdots & \vdots & \vdots\\ 0 & 0 & \dots & d_n\end{bmatrix} \tag{14}$$

where $d_i=\sum_{j=1}^{n}\omega_{ij}$.

Let $\boldsymbol{L}$ be the Laplacian matrix:

$$\boldsymbol{L}=\boldsymbol{D}_d-\boldsymbol{W} \tag{15}$$

Thus, $\boldsymbol{L}$ is a symmetric positive semidefinite matrix and its eigenvalues are λ_i. The eigenvalues arrange as follows:

$$0=\lambda_1\leq\lambda_2\leq\cdots\leq\lambda_n \tag{16}$$

For any vector $\boldsymbol{f}=(f_1, f_2, ..., f_i)$, there is:

$$\boldsymbol{f}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{f}=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\omega_{ij}(f_i-f_j)^2 \tag{17}$$

The indication vector is defined as $\boldsymbol{h}_j$:

$$\boldsymbol{h}_j=\begin{bmatrix} h_{1j} & h_{2j} & \dots & h_{nj}\end{bmatrix}^{\mathrm{T}} \tag{18}$$

$$h_{ij}=\begin{cases}\dfrac{1}{\sqrt{|A_j|}} & v_i\in A_j, i=1,2,\dots,n; j=1,2,\dots,m\\ 0 & v_i\notin A_j\end{cases} \tag{19}$$

Let $\boldsymbol{H}\in\mathbf{R}^{n\times m}$ be a matrix containing m indicator vectors as column vectors. The column vectors of $\boldsymbol{H}$ are orthogonal to each other, i.e., $\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}=\boldsymbol{I}$.

$$\boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{L}\boldsymbol{h}_i=\frac{1}{2}\sum_{u}\sum_{n}\omega_{un}(h_{iu}-h_{in})^2=$$
$$\frac{1}{2}\left[\sum_{u\in A_i,n\notin A_i}\omega_{un}\left(\frac{1}{\sqrt{A_i}}-0\right)^2+\sum_{u\in A_i,n\notin A_i}\omega_{un}\left(0-\frac{1}{\sqrt{A_i}}\right)^2\right]=$$
$$\frac{1}{2}\left[Cut(A_i,\bar{A}_i)\frac{1}{|A_i|}+Cut(\bar{A}_i,A_i)\frac{1}{|\bar{A}_i|}\right]=\frac{Cut(A_i,\bar{A}_i)}{|A_i|} \tag{20}$$

Equation (20) shows that for a subgraph A_i, its cut corresponds to $\boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{L}\boldsymbol{h}_i$. For m subgraphs, we can obtain:

$$RCut(A_1,A_2,\cdots,A_m)=\sum_{i=1}^{m}\boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{L}\boldsymbol{h}_i=\sum_{i=1}^{m}(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{H})_{ii}=Tr(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{H}) \tag{21}$$

The objective function is converted to:

$$\begin{cases}\min\limits_{\boldsymbol{H}\in\mathbf{R}^{n\times k}} Tr(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{H})\\ \text{s.t. } \boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}=\boldsymbol{I}\end{cases} \tag{22}$$

According to the Rayleigh quotient property [29], the minimum value of (22) is equal to the sum of the m smallest eigenvalues of $\boldsymbol{L}$. Finally, K-means clustering is performed on the matrix $\boldsymbol{F}$ composed of the eigenvectors corresponding to the minimum m eigenvalues of $\boldsymbol{L}$. Thereby, the clustering of normal data and bad data is realized, as shown in Fig. 6, and the normal data and bad data are completely separated. The flowchart and algorithm are shown in Fig. 7.

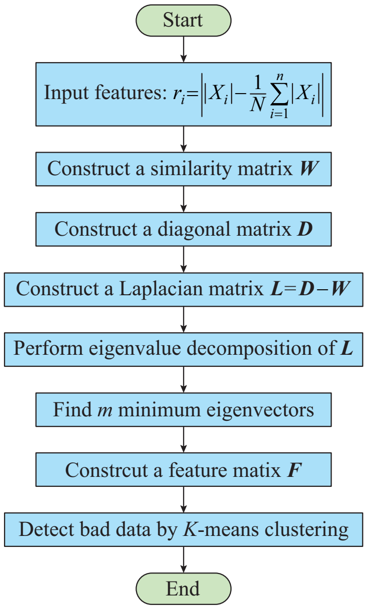

Start
Input features: $r_i=\left|\left|X_i\right|-\frac{1}{N}\sum_{i=1}^{n}\left|X_i\right|\right|$
Construct a similarity matrix $\boldsymbol{W}$
Construct a diagonal matrix $\boldsymbol{D}$
Construct a Laplacian matrix $\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W}$
Perform eigenvalue decomposition of $\boldsymbol{L}$
Find m minimum eigenvectors
Constrcut a feature matix $\boldsymbol{F}$
Detect bad data by K-means clustering
End

Fig. 7. Flowchart of spectral clustering.

According to the above process, the clustering features are first calculated as input by using the amplitude data, and then the similarity, diagonal and Laplacian matrices are constructed. Then decompose the eigenvalues of the matrix $\boldsymbol{L}$ to find the smallest m eigenvectors. The matrix $\boldsymbol{F}$ is composed of m eigenvectors. Clusters C_1 and C_2 can be obtained by K-means. C_1 contains normal data, and C_2 contains bad data. Hence, the bad data detection is realized.

B. Bad Data Detection Algorithm

The overall flow of the proposed algorithm is illustrated in Fig. 8. The algorithm has two parts. The first part is the event data and bad data identification method based on the decision tree. The details of the decision tree can be viewed in Fig. 3. The second part is the bad data detection method based on spectral clustering.

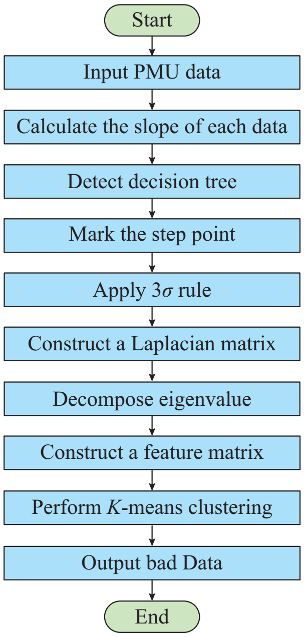

Start
Input PMU data
Calculate the slope of each data
Detect decision tree
Mark the step point
Apply 3σ rule
Construct a Laplacian matrix
Decompose eigenvalue
Construct a feature matrix
Perform K-means clustering
Output bad Data
End

Fig. 8. Flowchart of bad data detection algorithm.

IV. Case Studies

The algorithms presented in this paper are tested by simulation. In addition, field recorded PMU data is used to verify the method. The results are then compared with ensemble method (EM) [22] and the DBSCAN method [23].

A. Simulation

1) Simulation of Data Identification Method

When the power system is under the condition of normal operation, the data is ambient and has no external interference. The general expression of its signal is as follows:

$$x(t)=\sqrt{2}\,X_m\cos(2\pi f_0 t+\varphi_0)+n(t) \tag{23}$$

where X_m is the phasor amplitude, X_m = 57.73 V; f_0 is power frequency, $f_0=50$ Hz; φ_0 is the initial phases, $\varphi_0=0$; and the signal-noise ratio of $n(t)$ is 60 dB.

Different values of amplitude step (1 V, 1.5 V, 2 V, 3 V) are set with different durations (0.1 s, 0.5 s, 1 s, 2 s) to test the validity of the proposed method. The experiments are repeated 20 times. Also, many experiments with multiple parameters for the support vector machine (SVM) and the back-propagation algorithm (BP) have been conducted, and the best accuracy is used for comparison. The kernel function of SVM is on radial basis, where the gamma is 0.25, and the penalty factor is 10. The BP neural network has 3-layer, where the input layer has 4 nodes, the hidden layer has 12 nodes, and the output layer has 2 nodes. The number of iterations is 100. It is found that the EM and DBSCAN method cannot identify the step point and the average results of different methods are provided in Fig. 9.

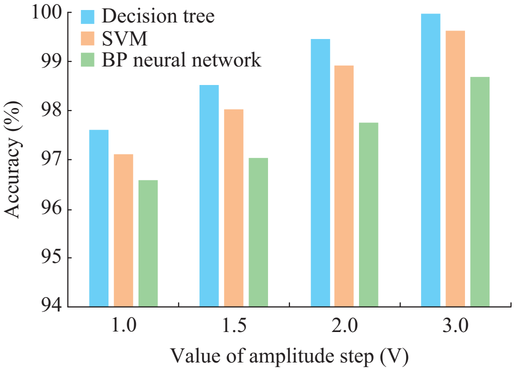


Fig. 9. Identification result of ambient data with different methods.

In Fig. 9, the horizontal axis represents the values of the amplitude step, and the vertical axis represents the average identification accuracy with different durations. As illustrated in the bar chart, the proposed method is more accurate than SVM and BP neural networks in different tests. When the amplitude increases, the identification accuracy increases, because the greater the step value is, the more obvious the features are.

The signal with amplitude and phase angle modulation is used to express the oscillation with low oscillation frequency, which can be expressed as:

$$\begin{aligned} x(t)=\sqrt{2}\,(X_m+X_d\cos(2\pi f_a t+\varphi_a))\cdot \\ \cos(2\pi f_0 t+X_k\cos(2\pi f_a t+\varphi_a)+\varphi_0)+n(t) \end{aligned} \tag{24}$$

where X_d is amplitude modulation depth, X_d = 0.5%; X_k is phase angle modulation depth, X_k = 5.7°; f_a is modulation frequency, f_a = 5 Hz; and φ_a is the initial phase angle of modulation part.

The identification accuracy of the event data and the oscillation data is then tested through the above steps. The results are provided in Fig. 10.

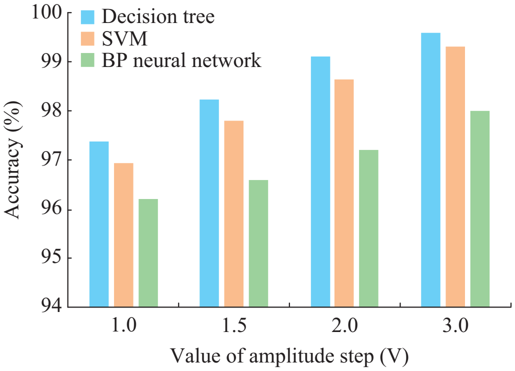


Fig. 10. Identification results of disturbance data with different methods.

Figure 10 shows that during the oscillation, event data can also be identified by the proposed method. The average accuracy of the proposed method is 98.5%, higher than the other two methods. The results of SVM are better than those of the BP neural network. In this paper, the sample data is small, thus BP neural network has no advantages. The accuracy of SVM is related to kernel function and other parameters [30]. Thus, the proposed method is more suitable.

2) Simulation of Data Detection Method

For the signal in (23), a number of single or contiguous bad data are artificially set, and the deviation range is from 0.3% to 5%. The detection results of the bad data with three methods is shown in Fig. 11.

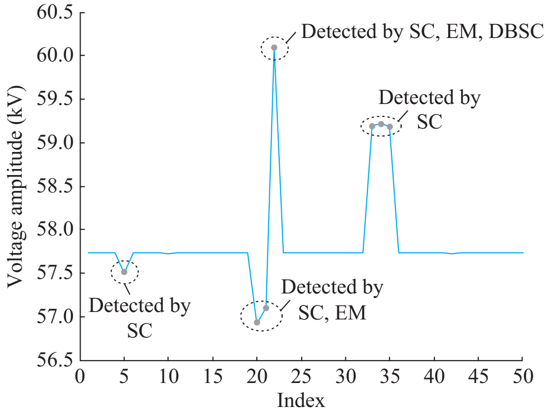


Fig. 11. Detection results of ambient data with different methods.

Figure 11 shows the ambient data with some bad data, which is marked by grey. In this figure, SC refers to the method proposed in this paper, EM is presented in detail in [22], and DBSCAN is the method proposed in [23]. It can be observed that the EM and DBSCAN method can effectively identify bad data when the amplitude changes greatly. However, the DBSCAN method cannot identify bad data with an amplitude that is close to normal value. When the deviation of bad data is small, it is closely related to the normal data. The bad data can easily be considered normal based on the strong density relationship. The EM also struggles to identify contiguous bad data as it is based on the prerequisite that the amplitudes of continuous bad data contrast dramatically. The proposed method can detect both single and contiguous bad data.

Using the signal in (23), the detection range of three methods by changing the deviation value of single bad data is compared. The results are provided in Fig. 12.

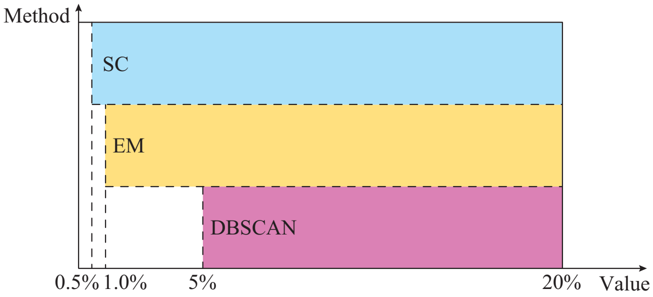


Fig. 12. Detection range of three methods with different deviation values.

Figure 12 illustrates that when the deviation value of bad data is lower than 1%, the EM cannot detect it. When the deviation value of bad data is lower than 5%, the DBSACN method cannot detect it. The proposed method can detect bad data with a deviation value from 0.5% to 20%.

Moreover, the ratio and position of bad data in Fig. 11 are randomly changed. The comparison of the detection ability of the three methods is as follows.

Table II shows that when the ratio of bad data is higher than 10%, the EM and DBSCAN method cannot detect them completely, while the proposed method can detect the ratio from 1% to 15%.

TABLE II
DETECTION CAPABILITY COMPARISON OF THREE METHODS WITH DIFFERENT RATIOS OF BAD DATA

Method	Detection capability				
	1.0%	2.5%	5.0%	10%	15%
SC	√	√	√	√	√
EM	√	√	√	×	×
DBSCAN	√	√	√	×	×

The signal in (24) represents the disturbance occurring in the system in which bad data is randomly set. The results of the three detection methods are provided in Fig. 13.

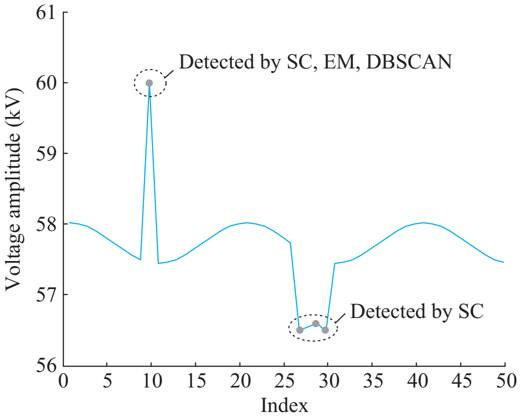


Fig. 13. Detection results of disturbance data with three methods.

Figure 13 indicates that when there is the data during a disturbance, the data divergent from the normal value can be detected by the three methods. However, the data with an amplitude close to the normal value cannot be detected by the DBSCAN method. Due to the close value, the density relationship is strong and can easily be classified into one cluster. EM cannot detect contiguous bad data because if the data in the middle is not much different from before and after, it is considered as normal. The proposed method can detect both single and contiguous bad data.

B. Field Data Verification

A PMU device suitable for a distribution network has been successfully developed in the laboratory. The PMU can measure the related parameters of the fundamental frequency, harmonics, and inter-harmonics in a distribution network in real time. Due to the synchronization signal loss on June 17th, 2019, the phase angle measurement jumps, which leads to the fluctuation of frequency and the change rate of frequency. Since the amplitude is corrected according to the frequency, the amplitude also jumps.

1) Field Data Verification of Data Identification Method

Aiming to verify the rationality of the parameter selection, the measurement of six other independent PMUs is validated to test the identification method. The results are provided in Fig. 14.

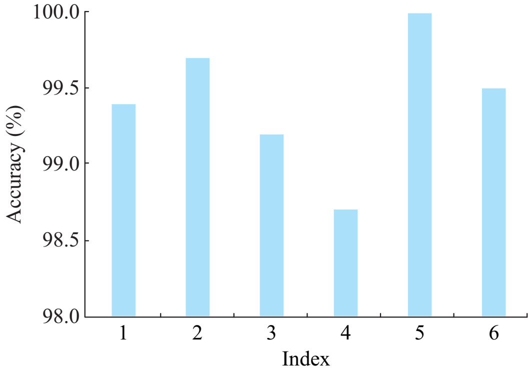


Fig. 14. Identification results of event data.

As seen in Fig. 14, the selected parameters of the decision tree are appropriate to the field data from other PMUs, and the accuracy of the event data identification method is higher than 98.6%.

2) Field Data Verification of Data Detection Method

Field data in the distribution network with event data and bad data is used to verify the algorithm. The comparison results are shown in Fig. 15.

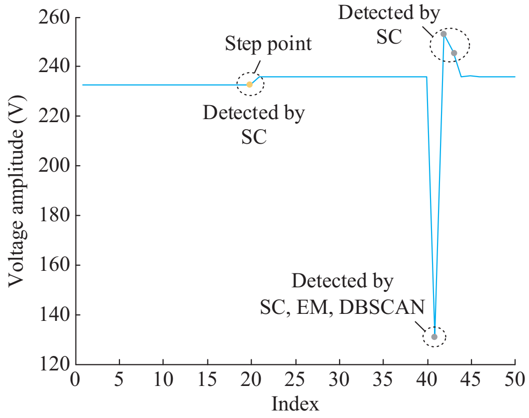


Fig. 15. Detection results of field data with three methods in a distribution network.

Figure 15 shows that the proposed method can identify the step point while the other two methods may decipher it as bad data. When the amplitude of bad data changes dramatically, all three methods can detect it. If the amplitude of bad data is close to the normal value, the DBSCAN method cannot detect it as the density relationship between bad data and normal data is strong. The EM cannot detect bad data with an amplitude that is near normal as the deviation does not exceed the set threshold. The proposed method can detect contiguous bad data regardless of the size of the deviation.

In addition, the bad data is artificially set for field data from a certain sub-synchronous oscillation in areas with renewable energy sources in western China. The filed data is in Fig. 16. The detection results are compared in Fig. 17.

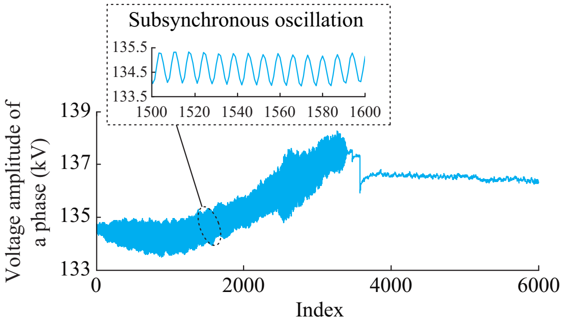


Fig. 16. PMU field data in a renewable energy area.

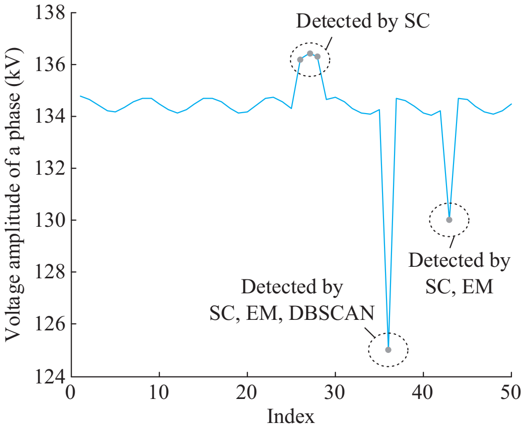


Fig. 17. Detection results of field data with three methods in a renewable energy area.

It can be seen from Fig. 16 that there is a single PMU measurement and the reporting rate is 100 Hz. The sub-synchronous oscillation lasts for about 33 s. Figure 17 shows that when there are contiguous bad data, the detection result of the proposed method is better than the other two methods. The detection range of the EM is larger than that of the DBSCAN method. Therefore, the proposed method can be practically applied to detect all the bad data in Fig. 17.

3) Performance Comparison of Different Methods

The running time of three detection methods is compared with different time windows. It should be pointed out that the running time of the proposed method in this paper does not include the bad data or event data identification process. When the time window contains 200 data points, the running time of the identification method is about 0.002 s. The calculation speed of the identification method is fast. The results of the running time of three detection methods are shown in Table III. It reports that the running time of the three methods increases as the time window expands. The EM runs longer than the other two methods because this method is more complicated. The running time of the DB-

SCAN method is close to that of the proposed method, as they are relatively simple and both belong to clustering methods.

TABLE III
COMPARISON OF RUNNING TIME OF THREE METHODS

Time window (s)	Data points	t_{EM} (s)	t_{DB} (s)	t_{SC} (s)
0.5	25	0.0052	0.0025	0.0028
1.0	50	0.0110	0.0049	0.0053
2.0	100	0.0260	0.0130	0.0110
3.0	150	0.0420	0.0210	0.0190
4.0	200	0.0570	0.0280	0.0260

When the ration of bad data is constant, the accuracy of the three methods is compared by changing the deviation range of bad data.

Table IV shows that the proposed method has higher accuracy than the other two methods in different scenarios. As the ratio of bad data increases, the detection accuracy of the three methods decreases. The accuracy increases as the deviation range of bad data increases. The DBSCAN method is more affected by the ratio and range.

TABLE IV
COMPARISON OF ACCURACY OF THREE METHODS

Ratio (%)	Range (%)	Type	Accuracy (%)
1.0	5.0	SC	99.61
		EM	98.48
		DBSCAN	93.15
1.0	10.0	SC	99.85
		EM	99.13
		DBSCAN	94.21
2.5	5.0	SC	98.53
		EM	96.32
		DBSCAN	90.54
2.5	10.0	SC	98.65
		EM	97.69
		DBSCAN	91.10
5.0	5.0	SC	96.48
		EM	93.52
		DBSCAN	86.36
5.0	10.0	SC	97.57
		EM	95.78
		DBSCAN	88.05

V. CONCLUSION

This paper proposes a data-driven PMU bad data detection algorithm. It only relies on a single PMU and does not need the system topology or parameters. It can improve the quality of PMU data, and lay a foundation for better application of PMU data to power systems. The main conclusions are as follows:

1) A data identification method based on a decision tree is proposed. Compared with the existing methods, it avoids mistaking event data for bad data by learning the slopes of each data.

2) A bad data detection method based on spectral clustering is developed. It can use the degree of association to cluster bad data. It detects bad data with small deviation values which is not easy to detect with the existing methods .

3) The simulation and field data tests prove that the proposed algorithm has effectiveness on bad data identification and detection. It can provide PMU data with high quality for the power systems.

This paper does not consider the bad data caused by PMU algorithms, which may cause long-term detection of bad data. Future works will focus on this problem.

REFERENCES

[1] B. L. R. Stojkoska and K. V. Trivodaliev, "A review of Internet of Things for smart home: challenges and solutions," *Journal of Cleaner Production*, vol. 140, pp. 1454-1464, Jan. 2017.

[2] K. Jia, Z. Xuan, T. Feng *et al*., "Transient high-frequency impedance comparison-based protection for flexible DC distribution systems," *IEEE Transactions on Smart Grid*, vol. 11, no. 1, pp. 323-333, Jan. 2020.

[3] J. D. L. Ree, V. Centeno, J. S. Thorp *et al*., "Synchronized phasor measurement applications in power systems," *IEEE Transactions on Smart Grid*, vol. 1, no. 1, pp. 20-27, Jun. 2010.

[4] M. U. Usman and M. O. Faruque, "Applications of synchrophasor technologies in power systems," *Journal of Modern Power Systems and Clean Energy*, vol. 7, no. 2, pp. 211-226, Mar. 2019.

[5] K. Jia, B. Yang, X. Dong *et al*., "Sparse voltage measurement-based fault location using intelligent electronic devices," *IEEE Transactions on Smart Grid*, vol. 11, no. 1, pp. 48-60, Jan. 2020.

[6] H. Liu, T. Bi, X. Chiang *et al*., "Impacts of subsynchronous and supersynchronous frequency components on synchrophasor measurements," *Journal of Modern Power Systems and Clean Energy*, vol. 4, no. 3, pp. 362-369, Jul. 2016.

[7] A. G. Phadke and T. Bi, "Phasor measurement units, WAMS, and their applications in protection and control of power systems," *Journal of Modern Power Systems and Clean Energy*, vol. 6, no. 4, pp. 619-629, Jul. 2018.

[8] A. Silverstein, "Synchrophasors & the grid," NASPI, Arlington, USA, Technical Report PNNL-SA-128949, Sept. 2017.

[9] A. Sundararajan, T. Khan, A. Moghadasi *et al*., "Survey on synchrophasor data quality and cybersecurity challenges, and evaluation of their interdependencies," *Journal of Modern Power Systems and Clean Energy*, vol. 7, no. 3, pp. 449-467, May 2019.

[10] California ISO, "Five year synchrophasor plan," California ISO, Folsom, USA, Technical Report, Nov. 2011.

[11] W. Qi, "Comparison of differences between SCADA and WAMS realtime data in dispatch center," in *Proceedings of the 12th International Workshop on Electric Power Control Centers*, Bedford Springs, USA, Jun. 2013, pp. 2-5.

[12] M. H. Amini, M. Rahmani, K. G. Boroojeni *et al*., "Sparsity-based error detection in DC power flow state estimation," in *Proceedings of 2016 IEEE International Conference on Electro Information Technology (EIT)*, Grand Forks, USA, May 2016, pp. 263-268.

[13] S. G. Ghiocel, J. Chow, G. Strefopoulos *et al*., "Phasor-measurement-based state estimation for synchrophasor data quality improvement and power transfer interface monitoring," *IEEE Transactions on Power Systems*, vol. 29, no. 2, pp. 881-888, Mar. 2014.

[14] K. D. Jones, A. Pal, and J. S. Thorp, "Methodology for performing synchrophasor data conditioning and validation," *IEEE Transactions on Power Systems*, vol. 30, no. 3, pp. 1121-1130, May 2015.

[15] N. Živković and A. T. Sarić, "Detection of false data injection attacks using unscented Kalman filter," *Journal of Modern Power Systems and Clean Energy*, vol. 6, no. 5, pp. 847-859, Sept. 2018.

[16] Y. Wu, Y. Xiao, F. Hohn *et al*., "Bad data detection using linear WLS and sampled values in digital substations," *IEEE Transactions on Power Delivery*, vol. 33, no. 1, pp. 150-157, Feb. 2018.

[17] J. Zhao, G. Zhang, M. L. Scala *et al*., "Enhanced robustness of state estimator to bad data processing through multi-innovation analysis," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 4, pp. 1610-1619, Aug. 2017.

[18] Q. F. Zhang and V. M. Venkatasubramanian, "Synchrophasor time

skew: formulation, detection and correction," in *Proceedings of 2014 North American Power Symposium (NAPS)*, Pullman, USA, Sept. 2014, pp. 1-6.

[19] M. Wu and L. Xie, "Online detection of low-quality synchrophasor measurements: a data-driven approach," *IEEE Transactions on Power Systems*, vol. 32, no. 4, pp. 2817-2827, Jul. 2017.

[20] M. Mardani, G. Mateos, and G. B. Giannakis, "Dynamic anomalography: tracking network anomalies via sparsity and low rank," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 1, pp. 50-66, Feb. 2013.

[21] K. Mahapatra, N. R. Chaudhuri, and R. Kavasseri, "Bad data detection in PMU measurements using principal component analysis," in *Proceedings of 2016 North American Power Symposium (NAPS)*, Denver, USA, Sept. 2016, pp. 1-6.

[22] M. Zhou, Y. Wang, A. K. Srivastava *et al*., "Ensemble-based algorithm for synchrophasor data anomaly detection," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 2979-2988, May. 2019.

[23] R. Vallakati, A. Mukherjee, and P. Ranganathan, "A density based clustering scheme for situational awareness in a smart-grid," in *Proceedings of 2015 IEEE International Conference on Electro/Information Technology (EIT)*, Dekalb, USA, May 2015, pp. 346-350.

[24] M. Esmalifalak, L. Liu, N. Nguyen *et al*., "Detecting stealthy false data injection using machine learning in smart grid," *IEEE Systems Journal*, vol. 11, no. 3, pp. 1644-1652, Sept. 2017.

[25] C. Huang, F. Li, D. Zhou *et al*., "Data quality issues for synchrophasor applications Part I: a review," *Journal of Modern Power Systems and Clean Energy*, vol. 4, no. 3, pp. 342-352, Jul. 2016.

[26] Z. Zhou and Y. Jiang, "NeC4.5: neural ensemble based C4.5," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 6, pp. 770-773, Jun. 2004.

[27] F. Pukelsheim, "The three sigma rule," *The American Statistician*, vol. 48, no. 2, pp. 88-91, May 1994.

[28] S. Mehrkanoon, C. Alzate, R. Mall *et al*., "Multiclass semisupervised learning based upon kernel spectral clustering," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 4, pp. 720-733, Apr. 2015.

[29] E. Louidor and B. H. Marcus, "Improved lower bounds on capacities of symmetric 2D constraints using Rayleigh quotients," *IEEE Transactions on Information Theory*, vol. 56, no. 4, pp. 1624-1639, Apr. 2010.

[30] F. Angiulli and A. Astorino, "Scaling up support vector machines using nearest neighbor condensation," *IEEE Transactions on Neural Networks*, vol. 21, no. 2, pp. 351-357, Feb. 2010.

Zhiwei Yang received the B.S. degree at School of Electrical and Electronic Engineering from North China Electric Power University (Baoding), Baoding, China, in 2015. He is currently working toward the Ph.D. degree with North China Electric Power University, Beijing, China. His research interests include PMU data quality assessment and correction.

Hao Liu received his Ph.D. degree at North China Electric Power University, Beijing, China, in 2015. He is currently an associate professor at North China Electric Power University, Beijing, China. His research interests include synchronized phasor measurement technique and its application.

Tianshu Bi received her Ph.D. degree at the Department of Electrical and Electronic Engineering in the University of Hong Kong, Hong Kong, China, in 2002. She is currently a professor at North China Electric Power University, Beijng, China. Her research interests include power system protection and control, synchronized phasor measurement technique and its application and fault diagnosis.

Qixun Yang received his Ph.D. degree at The University of New South Wales, Sydney, Australia, in 1982. He is a Chinese academician of engineering and a Professor at North China Electric Power University, Beijing, China. His research interests include power system protection and control, and substation automation.