IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received May 21, 2018, accepted June 28, 2018, date of publication July 5, 2018, date of current version August 7, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2852759

Outlier Data Treatment Methods Toward Smart Grid Applications

LI SUN[1,2], KAILE ZHOU[1,2,3], XIAOLING ZHANG[3], AND SHANLIN YANG[1,2]

[1]School of Management, Hefei University of Technology, Hefei 230009, China
[2]Key Laboratory of Process Optimization and Intelligent Decision-Making, Ministry of Education, Hefei University of Technology, Hefei 230009, China
[3]Department of Public Policy, City University of Hong Kong, Hong Kong

Corresponding authors: Kaile Zhou (kailezhou@gmail.com) and Xiaoling Zhang (xiaoling.zhang@cityu.edu.hk)

This work was supported in part by the National Natural Science Foundation of China under Grant 71501056, in part by the Anhui Science and Technology Major Project under Grant 17030901024, in part by the Hong Kong Scholars Program under Grant 2017-167, in part by the Research Grants Council of the Hong Kong Special Administrative Region, China, under Project CityU 11271716 and Project CityU 21209715, and in part by the China Postdoctoral Science Foundation under Grant 2017M612072.

ABSTRACT In a smart grid environment, advanced metering infrastructure (AMI) and intelligent sensors have been deployed extensively. As a result, large-scale and fine-grained smart grid data are more convenient to be collected, in which outliers exist pervasively, caused by system failures, environmental effects, and human interventions. Outlier deletion is always implemented in data preprocessing for improving data quality. However, due to the fact that real records that reflect rare and unusual patterns are also recognized as outliers, outlier mining is necessary to be carried out with the aim of discovering knowledge on abnormal patterns in power generation, transmission, distribution, transformation, and consumption. To the best of our knowledge, a comprehensive and systematic review of outlier data treatment methods is still lacked in the smart grid environment. We, in this paper, aim at presenting the review of outlier data treatment methods toward smart grid applications and categorize them into outlier rejection and outlier mining groups. Since we do this survey from the perspective of data-driven analytics and data mining methods, information security technologies are barely discussed in this paper. Based on a general overview of outlier data treatment methods, we make the contribution of providing the application scenarios of outlier rejection and outlier mining in the smart grid environment. With the construction of smart grid throughout the world, dealing with outlier data has become more crucial for the security and reliability of power system operation. Therefore, we also discuss some future challenges of outlier data treatment toward smart energy management.

INDEX TERMS Outlier data treatment, outlier rejection, outlier mining, smart grid, data preprocessing.

I. INTRODUCTION

Traditional energy systems are becoming more and more intelligent as they continuously integrate with emerging information technologies [1]. Worldwide deployment and construction of smart energy systems are being accelerated [2]. In smart grid environment, the wide deployment of advanced metering infrastructure (AMI) [3] makes it more convenient and easier to obtain massive and detailed smart grid data. Big data in smart grid are increasingly regarded as important strategic resources considering their potential business values [4], [5]. Besides, data driven analytics is always important for efficient and optimal operation of smart grid systems [6]–[9], especially for power supply demand balance, power supply reliability and state estimation. Zhou *et al.* [10] discovered household electricity demands based on a fuzzy clustering-based model. The discovered electricity demands of typical household groups could support production planning, thus to contribute to supply demand balance. Faza [11] used particle swarm optimization (PSO) algorithm to determine the optimal placement of photovoltaic (PV) sources with the objective of maximizing system reliability. Rahman and Venayagamoorthy [12] used genetic algorithm (GA) to improve the result of the proposed cellular computational network framework, and applied the hybrid estimator in state estimation of large power systems.

However, outlier data always exist in the smart grid data. Outlier data are the abnormal values that do not consist with the overall data distribution. Outlier data like noise reduce data quality and have adverse effects on the performance of data-based models [13], [14], thus to be regarded as "bad data". For the purpose of improving data quality, outlier rejection is always carried out in the process of data preprocessing. Furthermore, according to Hawkins's definition [15], "An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism." As Hawkins described, outlier data can be real collected records instead of bad data, and such real collected records always indicate significant abnormal patterns. Analyses on outliers with the aim of discovering valuable knowledge of rare but unusual patterns are known as outlier mining. Outlier mining is an interesting and important task in data mining. It has expansive applications in network intrusion detection [16], financial fraud detection [17], traffic anomaly detection [18], and price trend prediction in stock market [19]. Especially for smart grid applications, abnormal cases such as electric larceny [20] and equipment failure [21] can be discovered through outlier mining.

In the smart grid environment, previous research works mainly focused on data quality improvement, where outlier data were regarded as bad data and the purpose was to eliminate interferences in the built model [13], [22], [23]. To the best of our knowledge, a comprehensive review of outlier data treatment, including both outlier rejection and outlier mining, is still absent toward smart grid implications.

In this paper, a general overview of outlier data treatment methods is provided, followed by the application scenarios of outlier rejection and outlier mining in smart grid environment. With the further development of smart grid, the scale of power system becomes larger and its complexity is increasing. More challenges are brought by complex data like multi-source heterogeneous data and large scale real time data. In this paper, future challenges of outlier data treatment toward smart energy management are also discussed.

The rest of this work is organized as follows. In Section II, we give the background. Then, a general overview of outlier data treatment methods is provided in Section III, and their application scenarios in smart grid environment are presented in Section IV. Section V proposes the future challenges of outlier data treatment toward smart grid applications. Section VI makes conclusions.

II. BACKGROUND

A. SMART GRID

From the perspective of power line communication (PLC), bidirectional flow of electric power and interactive information communication between power grid companies and consumers are realized in smart grid [24], as shown in Fig. 1. Load balancing and efficiency improvement are enhanced due to timely interaction between power supply side and demand side [25]. Smart grid is considered to be an ecosystem where various kinds of renewable energy sources are connected [26]. As shown in Fig. 2, smart buildings and smart homes are equipped with power generation facilities to produce electric power for themselves and share the redundant part [26]. With the extensive deployment of intelligent transmission and distribution networks, connections among smart grid units become more extensive and complex [27]. Various data are involved in smart grid data analyses.

Outlier data exist pervasively in electricity consumption data, asset management data, and external data. Fig. 3 gives the details of the mentioned three types of smart grid data. As shown in Fig. 3, large amount of electricity consumption data are generated by different kinds of customers in smart grid, mainly including residents, commercial enterprises, industrial enterprises, and industrial parks. In the process of power generation, asset management data are mainly

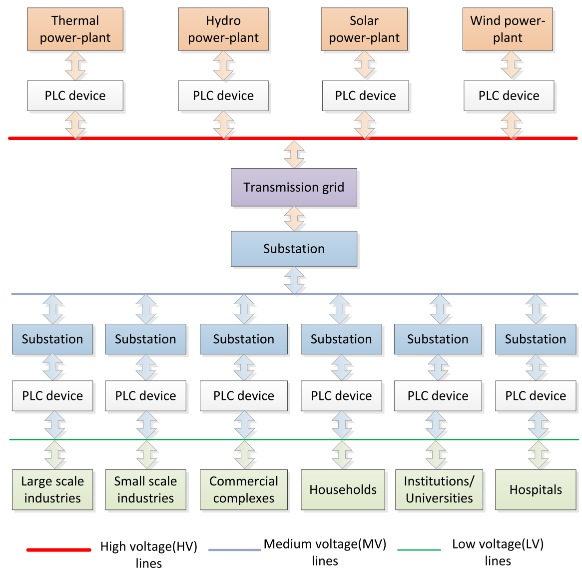


FIGURE 1. Smart grid power system architecture [24].

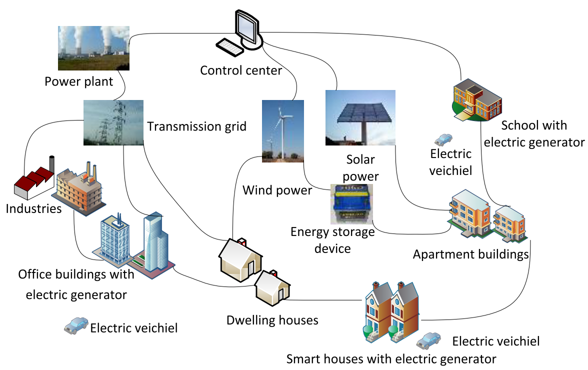


FIGURE 2. Smart grid ecosystem [26].

Electricity consumption data	Asset management data	External data
• Household electricity consumption data • Commercial electricity consumption data • Electricity consumption data of industrial enterprises • Electricity consumption data of industrial parks	• Equipment status data • Generator unit data • Transmission line data • Transformer substation data • Energy storage device data • Transaction data	• Meteorological data • GIS data • Smart building data • Electric vehicle data • Social media data

FIGURE 3. Smart grid data.

categorized into equipment operating data and transaction data. External data outside power systems are commonly applied to provide references and assist smart grid data analyses.

B. CAUSES OF OUTLIER DATA IN SMART GRID DATA

Complex and diverse outlier data are generated in the construction of automated, interactive electric power networks of smart grid. Major causes of outlier data are as follows.

(1) Data acquisition ability [28], [29]. The data acquisition devices such as smart meters and sensors have different performances in data acquisition frequency and accuracy. Measurement errors are usually caused by the limited capability of the devices. Besides, noise data can be generated when the anti-interference ability of the devices is weak.

(2) Failures in power systems. Many system failures, such as failures of data transmission system, faults of transmission equipment and power outage all can lead to the generation of outlier data [21], [30].

(3) Human factors. Activities in power systems such as hand off control, responding to contingencies and outage control are intervened by human [31]. Besides, human are also involved in the data collection process. Outlier data can be produced in these works because of human errors [32].

III. OUTLIER DATA TREATMENT METHODS

A. SVM-BASED METHODS

Support vector machine (SVM) learning method has shown prominent superiority in solving text classification and high dimension pattern recognition problems [33], [37]. Let n represent the size of the sample set $X = \{(x_1, y_1), \ldots, (x_i, y_i), \ldots, (x_n, y_n)\}$, where $i = 1, 2, \ldots, n$, $y_i \in \{+1, -1\}$, $x_i \in R^d$ and d denotes the dimension. SVM is designed to find the maximal margin hyperplane that classifies the samples into the two classes with label of $+1$ or -1. The described hyperplane can be denoted as $w \cdot x + b = 0$, where $w \in R^n$ and $b \in R$. w and b are estimated by minimizing the following objective function in (1), using training data.

$$\begin{aligned} &\min_{w} \frac{1}{2} \|w\|^2 \\ &\text{subject to } y_i \quad (w \cdot x_i + b) \geq 1, \quad i = 1, 2, \ldots, n \end{aligned} \tag{1}$$

However, the maximal margin hyperplane is very sensitive to outliers because it is determined by very few sample data. So considering the influence of outliers, a penalty factor C is introduced to develop (2). ξ_i means slack variable that expresses training error corresponding to x_i. It is calculated by input training samples and the user-specified constant C. Then, w and b are estimated by minimizing the following objective function in (2). Apparently, the introduced penalty factor C promotes SVM to become more tolerant to outliers and then it turns out to be the most commonly used method. Lin and Wang [14], [34] proposed a fuzzy SVM (FSVM) model to adjust the slack variable ξ_i of each training sample to reduce the impact of outliers. Actually, finding out how to reduce the sensitivity to outliers of SVM has already aroused widespread attention [35]–[39].

$$\begin{aligned} &\min_{w} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i \\ &\text{subject to } y_i \left(w^T x_i + b \right) \geq 1 - \xi_i \\ &\qquad \xi_i \geq 0, \quad i = 1, \ldots, n \end{aligned} \tag{2}$$

As we know, numerous dimensions in outlier mining can trigger the "curse of dimensionality", which refers to statistical significance problems that aroused by the sparse data. SVM has unique advantage towards high dimensional data [37]. The exploration of SVM based outlier mining methods were carried out. In 2009, Rajasegarar *et al.* [40] proposed an SVM based global outlier mining method. But the accuracy was not quite satisfying because of overlooking the spatial correlation of data. Zhang *et al.* [41] presented quarter-sphere based SVM outlier mining technique, and it was only applicable to spherical distributed data. Li [42] designed a spatiotemporal-attribute one-class hypersphere support vector machine, which improved the detection rate meanwhile reduced the false alarm rate and computational complexity. But the stability was badly affected by the number of attributes.

As an advanced machine learning method, SVM has been used in pattern recognition, classification, and regression analysis since it was proposed [43], [44]. SVM models have certain superiority for high dimension data and large scale data [45]. SVM models used to be sensitive to noise, but they were then improved and able to reach low outlier sensitivity. Further, SVM based outlier mining methods were developed.

B. PROXIMITY-BASED METHODS

Proximity-based methods ascertain outliers by defining rules of proximity measurement rather than building up models to fit data distribution. Specifically, proximity-based methods can be categorized into distance based methods and density based methods.

Distance based methods measure the proximity between objects through calculating distances. The lower value of the distance metric indicates the closer proximity. Ramaswamy *et al.* [46] suggested simply adopting the distance between object p and its kth nearest neighbor as the score of being an outlier and noted the score as $kNN(p)$, as shown in (3). k should be pre-determined by users. Then, the distance between object p and its kth nearest neighbor, $d_k(p)$, was calculated as the kNN score of being an outlier, commonly using Euclidean distance and Manhattan distance. Outliers shall be the objects that with highest kNN scores. Angiulli and Pizzuti [47] synthesized distances between object p and its k nearest neighbors. They were no longer just applying a single $d_k(p)$ to calculate the score of being an outlier, $agg\_kNN(p)$, as shown in (4). With a user pre-determined k, all the $d_i(p)$ (i from 1 to k) for an object p were added up to calculate the agg_kNN score of being an outlier. Outliers shall be the objects that with highest agg_kNN scores. Yu *et al*. [48] combined (3) and (4). They used the Local Isolation Coefficient of object p, which denoted as $LIC(p)$, to figure out the score of being an outlier. That is described in (5).

$$kNN(p) = d_k(p) \tag{3}$$

$$agg\_kNN(p) = \sum_{i=1}^{k} d_i(p) \tag{4}$$

$$LIC(p) = d_k(p) + \sum_{i=1}^{k} \frac{d_i(p)}{k} \tag{5}$$

The distance based methods were easy to suffer from local density of a dataset, so density based methods were developed. Breunig *et al*. [49] proposed local outlier factor (LOF) which estimated the density around p through the reachability density between object p and q. First, a $k - distance(p)$ of object p is defined as follows. Denote the distance between object p and q as $d(p, q)$, then $k - distance(p) = d(p, q)$ if:

1) $d(p, o) \leq d(p, q)$, for at least the number of k objects $o(p \neq q)$;
2) $d(p, o) < d(p, q)$, for at most the number of $k - 1$ objects o $(p \neq q)$.

Then, LOF is shown in detail in (8) where $N_k(p)$ represents the set of k nearest neighbors of object p, namely,

$$N_k(p) = \{q|d(p, q) \leq d_k(p), q \neq p\} \tag{6}$$

where $Ird_k(p)$ denotes local reachability density,

$$Ird_k(p) = \frac{|N_k(p)|}{\sum_{q \in N_k(p)} \max\{k - distance(q), d(p, q)\}} \tag{7}$$

Higher LOF score than 1 indicates the bigger possibility of being an outlier. Since then, it has become one of the most commonly used outlier mining methods [50]–[52]. Tang *et al*. [53] did not believe that low density was a necessary condition of being an outlier. So they put forward connectivity based outlier factor (COF). COF selects a set of nearest neighbors using a set-based shortest path [53]. The selected set is further applied to find the relative density of test points within average chain distance. When the outlier is in the middle of two clusters with similar density, COF behaves more effective. But when the outlier is between a sparse and a dense cluster, LOF and COF both have poor performance. Jin *et al*. [54] offered a new algorithm based on symmetric neighborhood relations named influenced local outlier factor (INFLO). The forward and reverse neighbors were both taken into account when evaluating the density distribution, thus to overcome the mentioned shortcoming of LOF and COF.

$$LOF(p) = \frac{1}{|N_k(p)|} \sum_{q \in N_k(p)} \frac{Ird_k(q)}{Ird_k(p)} \tag{8}$$

C. HYBRID METHODS

In order to enhance the efficiency and accuracy, models are usually combined in outlier data treatment. Nagi *et al*. [55] developed a hybrid GA-SVM model to combine Genetic Algorithm (GA) with SVM to discover outlier patterns. Fei and Zhang [56] used GA to select appropriate parameters for SVM in outlier mining. Higher accuracy were achieved in fault diagnosis. Yang *et al*. [57] conducted local outlier mining by associating clustering with distance based approaches. Firstly, they used hierarchical clustering algorithm and K-means algorithm to divide the dataset into several clusters. Then distance based outlier mining algorithm was employed to recognize local outliers in each cluster. Qian *et al*. [58] proposed an outlier mining algorithm based on genetic clustering. It gave full play to the local convergence of K-means algorithm, also, the global searching ability of genetic algorithm. Ping *et al*. [59] proposed PMLDOF algorithm based on multiple DBSCN clustering to prune data. PMLDOF was an improvement of the distance based algorithm LDOF. It could successfully select cluster edge points, meanwhile avoided the false shear of outliers.

In this section, we introduce several outlier data treatment methods, including SVM-based methods, proximity-based methods, and hybrid methods. Table 1 divides proximity-based methods into distance based and density based methods, and gives the advantages and disadvantages of each method. Now taking efficiency and accuracy both into account, explorations and improvements of outlier data treatment are still based on these approaches [52], [57]–[61].

TABLE 1. Comparisons among outlier data treatment methods.

Methods	Advantages	Disadvantages	Applications in smart grid
SVM based methods	Have superiority in high dimensional data [37]	Sensitive to outliers in small dataset [14]	[62, 63]
Distance based methods	Simple, using distance to identify outliers [46, 47]	Unable to find local outliers [49]	[64-66]
Density based methods	Able to find local outliers [49]	Difficult to deal with the sparse situation of high dimensional data [53, 54]	[67]
Hybrid methods	With improved efficiency and accuracy [55, 58]	Have increased complexity, usually have difficulties in pre-setting parameters [56, 57]	[55, 56, 68, 69]

IV. APPLICATION SCENARIOS OF OUTLIER DATA TREATMENT IN SMART GRID

A. RELATIONSHIP BETWEEN OUTLIER REJECTION AND OUTLIER MINING

In Fig. 4, the process model of outlier data treatment in smart grid is provided. As shown in Fig.4, outlier rejection takes place in the process of data cleaning for the

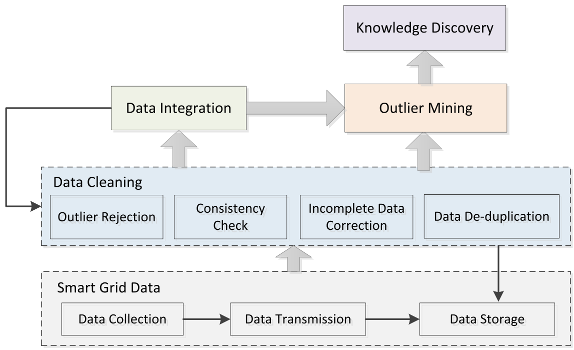


FIGURE 4. A process model of outlier data treatment in smart grid.

TABLE 2. Applications of outlier rejection in smart grid.

Application scenario	Methods	Achievements	Refs
Load forecasting	Distance based outlier rejection	Identified and immediately rejected global outliers	[22]
State estimation	Distance based outlier rejection	Identified outliers quickly, avoided pollution and residual submersion	[23]
Anomaly detection of power equipment	Hybrid method of big data analysis and unsupervised learning	Realized outlier detection and rejection in dynamic data with high accuracy rate	[75]
Outlier rejection in thermal power plants	Modified grubbs method	Simple and robust for identifying outliers	[76, 77]
Fault diagnosis	Modified Partial Robust M-regression (MPRM)	More tolerant to outliers than partial least squares (PLS) regression	[78]

purpose of improving data quality. Data cleaning is considered to be a repeated process which aims at continuously discovering data quality problems such as incomplete, inconsistency, duplicate data, and solving them [70], [71]. Besides, data preprocessing technologies including data cleaning approaches and data integration approaches are developed in order to improve the quality of data mining [72], [73]. Outlier mining are explored based on the preprocessed data and aimed at discovering knowledge in smart grid. In the following two Sections, outlier rejection and outlier mining scenarios in smart grid are provided in detail.

B. OUTLIER REJECTION

Outlier rejection is implemented in the process of data cleaning, in which case outlier data are regarded as bad data. The existing studies in smart grid environment explored outlier rejection from the following two perspectives. The first one focused on conducting outlier detection with the aim of removing them or realizing outlier correction. The second one aimed at developing models that were robust or insensitive to outliers. All these works have been concluded in Table 2. As shown in Table 2, except the aforementioned methods in Section III, there are some traditional statistical methods which are seldom used recently [76]–[78]. The statistical model based methods try to build the data distribution model from the complex real world and usually have limited accuracy.

C. OUTLIER MINING

Fig. 5 concretely provides the knowledge discovery scenarios in smart grid based on outlier mining. Next, outlier mining in electricity consumption data, asset management data and external data is presented respectively.

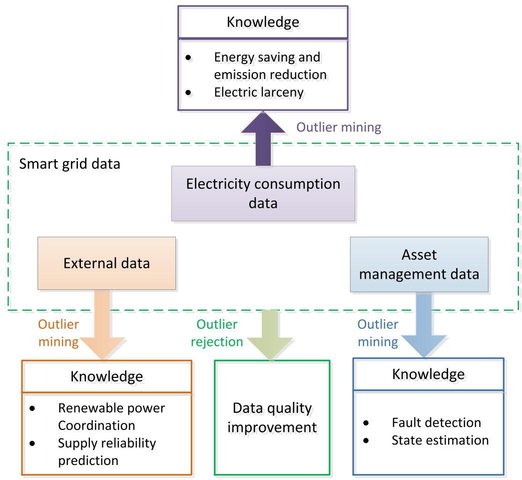


FIGURE 5. Applications of outlier mining in smart grid.

Demand side management [79] theory reveals that customers adjust their behaviors because of the changing electricity prices and incentives [80]–[82]. Customer groups with outlier electricity consumption patterns (e.g., high amount and fluctuation in the load curve) are recognized as potential customers of demand response (DR) projects for realizing energy conservation and emission reduction [10], [83], [84]. Besides, mining outlier consumption patterns during the special time period like Chinese Spring Festival is believed to support the production planning, thus to balance electricity supply demand [84]. Rush hour outage is of high probability to be avoided if outliers like extremely high consumptions are discovered and the corresponding strategies on load reduction or transmission are provided.

In addition, detailed consumption data provide innovative ideas for identifying electric larceny based on outlier mining [62], [68]. Traditional theft detection methods such as regular inspection, regular meter checking, and user reporting have low efficiency and poor accuracy. Nizar *et al.* [85] compared load data and time domain data in a feature extraction based non-technical losses detection method. They found that load data were more representative to describe consumption behaviors. Sheng *et al.* [20] held that the current electric larceny identification followed the following two kinds of ideas. One assumed that for a certain user, there would be obvious differences between an ordinary load curve and a curve with electric larceny. Based on that, electric larceny could be ascertained by extracting historical characteristics of the user. Another viewpoint focused on classifying customer groups. Then in each group, conducting comparisons among customers' load curves. From the former idea, Cheng *et al.* [66] applied distance based outlier mining for electric larceny detection. They believed that three-phase voltage and current unbalance rate was nearly a fixed value for normal customer's ordinary usage. So, if there exists a stealing, the reflected voltage and current abnormality would make the customer become a global outlier. Depuru *et al.* [63] focused on detecting electricity theft through data classification. They trained SVMs with historical data to identify abnormal electricity consumption patterns. As for the second perspective, dos Angelos *et al.* [65] used C-means based fuzzy clustering to group customers. Electric larcenies were consequently detected and identified according to a unitary index score.

Asset management data are mainly equipment data, which produced by instrumentations and sensor equipment in the process of power generation, transmission, distribution and substation. As we know, outlier data are usually bad data that do not consist with data distribution. Therefore, they have negative impacts on fault diagnosis [78] and state estimation [86]. However, outlier data can also be helpful to estimate the running state because of their exposure of abnormal changes in equipment operation. Jamali *et al.* [87] presented a new method to find the fault location by applying outlier identification technique, which did not require the fault type. Yan *et al.* [30] held that it was necessary to extract effective fault information from transmission equipment state data. Compared with delete outliers directly, it could avoid the loss of useful information. Yu *et al.* [88] believed the variations of harmonic data include normal variations and abnormal changes. The former were caused by load changes while the later aroused from equipment failures and acquisition errors. They identified outliers from harmonic currents and discovered abnormal changes in a timely manner, also, with rare mistake faults. Shen *et al.* [21] analyzed the relationship of transmission equipment's adverse conditions, failure modes and abnormal symptoms. They used bias causal network to conclude fault patterns that power transmission equipment may suffer from.

External data can affect the stability and reliability of power systems. Geographic Information System (GIS) data describe the location of devices or power grids. They play an important role in the selection of sites and the dispatch work. Besides, since power load is extremely sensitive to temperature and weather conditions, electric outages are frequently triggered by abnormal temperature or climate changes [89]. Kenward and Raja [90] pointed out that nearly 80% of large-scale outages were caused by abnormal severe weather between 2003 and 2012. Smart grid is prone to failures if affected by the abnormal weather. Storms and hurricanes usually cause failures and damages of overhead transmission lines. But as described in [91], such outliers can be applied to predict outage and locate fault area, speeding up the process of fault warning and recovery. In addition, considering that renewable energy generation is sensitive to climate changes [92], [93], outlier mining in the external data plays a vital role in coordinating renewable energy power generation. Hence, outlier mining in external data are really important for maintaining safe and stable operation of power systems.

V. FUTURE CHALLENGES OF OUTLIER DATA TREATMENT IN SMART GRID

The construction of smart grid makes traditional power systems gradually expose to big data problems [94]. Complex data such as multi-source heterogeneous data and large scale real-time data have brought great difficulties in outlier data treatment. Except that, outlier data visualization is also directed to severe challenges in smart grid.

A. INTEGRATION OF MULTI-SOURCE HETEROGENEOUS DATA

Now in the smart grid, data are obtained from hundreds of millions of smart meters, smart appliances, and distributed storage devices. Plus, different power companies or organizations adopt different definition, storage and management standards. So, the acquired multi-source data are usually heterogeneous and independent. Fig. 6 provides data integration scenarios in smart grid. Now, the integration requirements of Energy Management System (EMS), Distribution Management System (DMS), Energy Storage System (ESS), and other information systems are increasing. Based on that, the business integration is carried out [95]. Meter data management of AMI is integrating with other business management systems. Then, data sharing among automatic measurement systems, marketing management systems, and production scheduling systems is gradually accomplished.

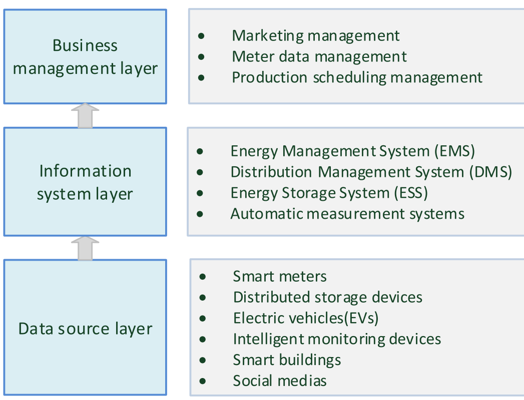


FIGURE 6. Data integrations in smart grid.

In the mentioned background of data integration in smart grid, multi-source heterogeneous data integration techniques become crucial.

Multi-source heterogeneous data in smart grid bring many challenges for outlier data treatment [96]. Spatial outliers widely exist in power systems. These outliers must be considered in the detection of abnormal sensors and abnormal space weather patterns. Janeja and Atluri [97] held that the existing spatial outlier mining methods basically focused on separated autocorrelation but ignored the heterogeneity among spatial objects. They proposed heterogeneous neighborhood spatial outlier mining method, but the time complexity was very large. Solaimani *et al*. [98] used statistical techniques to perform outlier mining on heterogeneous dataset. They integrated heterogeneous data streams through separating data collection, data preprocessing, and data analysis. However, it was not realistic to attach the ''global'' technology directly on heterogeneous data to discover all outliers. The results of data analysis need to be acquired in time, which was also a great challenge to processing efficiency.

Finally in this section, we summarize 3 key research issues on outlier data treatment for multi-source heterogeneous data in smart grid.

(1) Multi-format data integration. New technologies in smart grid are generally based on sensor networks and information systems. The integration of heterogeneous data sources is closely related to database structure [99]. Due to the lack of data description (e.g. primary keys, foreign keys, etc.), format conversion of heterogeneous data is along with information loss. This increases the difficulty of outlier pattern mining in power systems.

(2) Data interoperability among systems. Different data formats are hard to be supported at the same time by a certain system. So studies on the compatibility requirements and the translation toward common formats are crucial to outlier data treatment.

(3) Cooperation of multiple data centers. The integration of multi-source heterogeneous data tends to correlate with network server and local storage. As higher requirement on integration efficiency is expected in the smart grid environment, the development of information sharing platforms among multiple data centers is quite vital to efficiency enhancement.

B. REAL TIME PROCESSING OF LARGE-SCALE DATA

Large amounts of electricity production and consumption data are being generated, collected and stored in power systems. Smart meters usually collect consumption data every 15 minutes. For a utility company with AMI deployment, the amount of collected data dramatically increases from 24 million a year to 220 million a day [91]. With the exponential growth of smart grid data, large-scale scheduling problems [100] in smart grids become more severe.

Outlier data treatment requires high efficiency in massive amount of power data. Data processing needs to be completed in very short time. Moreover, supply-demand balance and instantaneous response are increasingly important for power systems, which requires large-scale data to be dealt with in real time [101]. This requirement is especially reflected in the monitoring of equipment, as well as the operation status of power grids. Besides, smart grids are exposed to many malicious attacks [6], making it important to deal with abnormal situations timely [74]. In recent years, distance based and density based outlier mining methods are often being discussed and improved. But they have obvious disadvantages on space and time complexity when applied to large datasets. The traditional statistical methods tried to use simple models to summarize complex situations of real data. Besides, the threshold values should be pre-set by human so that the detection accuracy was limited [102]. When these methods were used to deal with real-time outliers, they performed quite inefficiency [103]. Besides, the existing outlier mining methods, like decision trees, the optimal path of forest [104], fuzzy C means clustering [105] and kd-tree [52], were mostly offline methods. SVM was really limited to deal with real-time data because it required to pre classify all the normal and abnormal situations [106]. Neural networks were used in real-time mining, and performed well when there was only few outliers. But training data as well as threshold setting became two greatest difficulties that limited the better application [107].

Real-time data in smart grids has the characteristics of sequence uncontrollable and large scale. They are directed to many practical problems in outlier data treatment, which come down to the following 3 points.

(1) Uncertainties in dynamic data. The existing outlier mining methods built learning models by training history data. Then, used the established models as the basis to recognize outlier patterns. Due to the fact that distributions of actual stream data are dynamic, false alarms of outlier data are prone to appear. In [108], the influence of uncertainties on the performances of power system were studied in detail.

(2) Pre-set of parameters. Parameters of the complex algorithms need to be set up in advance. Therefore, the results will be directly affected if these parameters are not appropriately set [107], [109].

(3) Sparse data problems. Real time data or high frequency data are acquired with sparsity, which greatly increases the difficulty of getting valuable information from the large amount of data.

C. OUTLIER VISIUALIZAION

Compared with the traditional outlier mining methods, outlier visualization reflects human-computer interaction more friendly. The visualization technologies [110] can convert complex feature description data into images or graphics. The complexity of massive data is significantly reduced, which is conducive to the efficiency improving. With the help of visualization technologies, outliers can be quickly and accurately distinguished from normal states.

Plenty of traditional visualization methods were designed toward power system operation data. For node data, two-dimensional visualization methods like thermometer method [111] and histogram method were developed. As for branch operation data, a simple pie chart [111] was applied, in which diameter denoted power value and fan-shaped area expressed loading rate. But for the large network, these methods cannot give an overview of the abnormal areas' distribution. Now in power grid, transformations that from static to dynamic, from two-dimensional to three-dimensional have been realized in the visualization of real-time monitoring [112]–[114]. Location display of abnormal situations is achieved with the combination of pie chart, Gantt chart, radar chart, and trend chart [115]. In addition, critical abnormal alarm and warning information have combined with trend chart in power system. For equipment failure alarm, fault types are marked in the geographic map and fault location is displayed. Besides, it is worth noting that three-dimensional curves of power grid real-time monitoring have been developed. These three-dimensional curves contrast load situations of different regions at the current day and graphically display the trend and characteristics of power grid data. Thus, the understanding of abnormal situation is grasped.

Through data visualization, people are effectively liberated from the complex, massive data. Besides, more intuitive understandings of the power grid status are established in finding outlier situations. Outlier visualization in power systems has achieved good results. But many challenges still need to be concentrated in the smart grid construction background.

(1) Visualization of on-line data. In smart grid environment, the automatic visualization technologies of abnormal records are urgently required, while the dynamic property of real-time data brings difficulties in visualization efficiency and accuracy.

(2) Periodical variation of data. Using linear mapping techniques to visualize time series data can find data trends easily so that to find outliers. However, these methods are likely to ignore the periodicity of data.

(3) Environment building. Outlier visualization is meant to support the interactive analysis of complex anomalies. Therefore, the construction of collaborative environment that support data sharing is prerequisite for outlier visualization.

VI. CONCLUSIONS

Outlier data in power systems have become more complex and diverse in the context of fast-growing smart grid. Outlier data need to be properly dealt with, in order to better analyze electricity consumption data, asset management data, and external data. Although outlier data are usually bad data which reduce data quality and interfere with data analysis model, they can be unusual records that reflect true anomalies. In this paper, a comprehensive review of outlier data treatment in smart grid environment including outlier rejection and outlier mining is provided. With further construction of smart grid, the scale of power system becomes larger and the complexity continues to increase. Future challenges in outlier data treatment are brought by multi-source heterogeneous data, large scale real time data and outlier visualization, which are also discussed.

REFERENCES

[1] H. Jiang, K. Wang, Y. Wang, M. Gao, and Y. Zhang, "Energy big data: A survey," *IEEE Access*, vol. 4, pp. 3844–3861, 2016.

[2] X. Fang, S. Misra, G. Xue, and D. J. Yang, "Smart grid—The new and improved power grid: A survey," *IEEE Commun. Surveys Tuts.*, vol. 14, no. 4, pp. 944–980, 4th Quart., 2012.

[3] R. E. Brown, "Impact of smart grid on distribution system design," in *Proc. IEEE Power Energy Soc. General Meeting*, Jul. 2008, pp. 986–989.

[4] S. LaValle, E. Lesser, R. Shockley, M. S. Hopkins, and N. Kruschwitz, "Big data, analytics and the path from insights to value," *MIT Sloan Manage. Rev.*, vol. 52, no. 2, pp. 21–32, 2011.

[5] H. G. Miller and P. Mork, "From data to decisions: A value chain for big data," *IT Prof.*, vol. 15, no. 1, pp. 57–59, 2013.

[6] O. Kosut, L. Jia, R. J. Thomas, and L. Tong, "Malicious data attacks on the smart grid," *IEEE Trans. Smart Grid*, vol. 2, no. 4, pp. 645–658, Dec. 2011.

[7] Y. Yuan, Z. Li, and K. Ren, "Quantitative analysis of load redistribution attacks in power systems," *IEEE Trans. Parallel Distrib. Syst.*, vol. 23, no. 9, pp. 1731–1738, Sep. 2012.

[8] S. Ruj and A. Pal, "Analyzing cascading failures in smart grids under random and targeted attacks," in *Proc. IEEE 28th Int. Conf. Adv. Inf. Netw. Appl. (AINA)*, May 2014, pp. 226–233.

[9] A. Hamlyn, H. Cheung, T. Mander, L. Wang, C. Yang, and R. Cheung, "Network security management and authentication of actions for smart grids operations," in *Proc. IEEE Elect. Power Conf. (EPC)*, Oct. 2007, pp. 31–36.

[10] K. L. Zhou, S. Yang, and Z. Shao, "Household monthly electricity consumption pattern mining: A fuzzy clustering-based model and a case study," *J. Cleaner Prod.*, vol. 141, pp. 900–908, Jan. 2017.

[11] A. Faza, "A probabilistic model for estimating the effects of photovoltaic sources on the power systems reliability," *Rel. Eng. Syst. Saf.*, vol. 171, pp. 67–77, Mar. 2018.

[12] M. A. Rahman and G. K. Venayagamoorthy, "A hybrid method for power system state estimation using cellular computational network," *Eng. Appl. Artif. Intell.*, vol. 64, pp. 140–151, Sep. 2017.

[13] W. Chen, K. Zhou, S. Yang, and C. Wu, "Data quality of electricity consumption data in a smart grid environment," *Renew. Sustain. Energy Rev.*, vol. 75, pp. 98–105, Aug. 2017.

[14] C. F. Lin and S. D. Wang, "Training algorithms for fuzzy support vector machines with noisy data," *Pattern Recognit. Lett.*, vol. 25, no. 14, pp. 1647–1656, Oct. 2004.

[15] D. M. Hawkins, *Identification of Outliers*. New York, NY, USA: Springer, 1981, p. 860.

[16] J. Jabez and B. Muthukumar, "Intrusion detection system (IDS): Anomaly detection using outlier detection approach," *Procedia Comput. Sci.*, vol. 48, pp. 338–346, May 2015.

[17] M. Ahmed, A. N. Mahmood, and M. R. Islam, "A survey of anomaly detection techniques in financial domain," *Future Gener. Comput. Syst.*, vol. 55, pp. 278–288, Feb. 2016.

[18] J. Zhang, H. Li, Q. Gao, H. Wang, and Y. Luo, "Detecting anomalies from big network traffic data using an adaptive detection approach," *Inf. Sci.*, vol. 318, pp. 91–110, Oct. 2015.

[19] L. Zhao and L. Wang, "Price trend prediction of stock market using outlier data mining algorithm," in *Proc. IEEE 5th Int. Conf. Big Data Cloud Comput.*, Aug. 2015, pp. 93–98.

[20] S. Sheng, W. Changjiang, and M. Jun, "Attacker's perspective based analysis on cyber attack mode to cyber-physical system," *Power Syst. Technol.*, vol. 38, no. 11, pp. 3115–3120, 2014.

[21] X. Shen, M. Cao, W. Gao, X. Wang, and Q. Liu, "Research on dynamic diagnosis technology of power transmission and transformation equipment," *Chin. J. Electron Devices*, vol. 38, no. 5, pp. 175–1181, Oct. 2015.

[22] A. I. Saleh, A. H. Rabie, and K. M. Abo-Al-Ez, "A data mining based load forecasting strategy for smart electrical grids," *Adv. Eng. Inform.*, vol. 30, no. 3, pp. 422–448, Aug. 2016.

[23] D. L. Jiang, K. W. Wang, and X. D. Wang, "Clustering method of fuzzy equivalence matrix to bad-data detection and identification," *Power Syst. Protection Control*, vol. 39, no. 21, pp. 1–6, Nov. 2011.

[24] K. Sharma and L. M. Saini, "Power-line communications for smart grid: Progress, challenges, opportunities and status," *Renew. Sustain. Energy Rev.*, vol. 67, pp. 704–751, Jan. 2017.

[25] K. Zhou, S. Yang, and Z. Shao, "Energy Internet: The business perspective," *Appl. Energy*, vol. 178, pp. 212–222, Sep. 2016.

[26] Accenture. (2015). *Business Ecosystem Outlook of Chinese Energy Internet*. [Online]. Available: https://www.accenture.com/t00010101T000000__w__/cn-zh/_acnmedia/Accenture/Conversion-Assets/DotCom/Documents/Local/cn-zh/PDF/Accenture-Insight-Commercial-Ecology-Perspective-China-Energy-Internet.pdf#zoom=502015

[27] Y. Yan, Y. Qian, H. Sharif, and D. Tipper, "A survey on smart grid communication infrastructures: Motivations, requirements and challenges," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 1, pp. 5–20, 1st Quart., 2013.

[28] S. S. S. R. Depuru, L. F. Wang, and V. Devabhaktuni, "Smart meters for power grid: Challenges, issues, advantages and status," *Renew. Sustain. Energy Rev.*, vol. 15, no. 6, pp. 2736–2742, Aug. 2011.

[29] S. Park, H. Kim, H. Moon, J. Heo, and S. Yoon, "Concurrent simulation platform for energy-aware smart metering systems," *IEEE Trans. Consum. Electron.*, vol. 56, no. 3, pp. 1918–1926, Aug. 2010.

[30] Y. Yan, G. Sheng, Y. Chen, X. Jiang, Z. Guo, and S. Qin, "Cleaning method for big data of power transmission and transformation equipment state based on time sequence analysis," *Dianli Xitong Zidonghua/Autom. Electr. Power Syst.*, vol. 39, no. 7, pp. 138–144, 2015.

[31] C. Zhang *et al.*, "Human-centered automation for resilient nuclear power plant outage control," *Automat. Construction*, vol. 82, pp. 179–192, Oct. 2017.

[32] M. Last and A. Kandel, "Automated detection of outliers in real-world data," in *Proc. 2nd Int. Conf. Intell. Technol.*, 2001, pp. 292–301.

[33] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[34] C. F. Lin and S. D. Wang, "Fuzzy support vector machines," *IEEE Trans. Neural Netw.*, vol. 13, no. 2, pp. 464–471, Mar. 2002.

[35] Y. Wang, S. Wang, and K. K. Lai, "A new fuzzy support vector machine to evaluate credit risk," *IEEE Trans. Fuzzy Syst.*, vol. 13, no. 6, pp. 820–831, Dec. 2005.

[36] J. Sun and Y. Zhou, "Noise reduction of chaotic systems based on least squares support vector machines," in *Proc. IEEE Int. Conf. Commun., Circuits Syst.*, Jun. 2006, pp. 336–339.

[37] T. Y. Wang and H. M. Chiang, "Fuzzy support vector machine for multi-class text categorization," *Inf. Process. Manage.*, vol. 43, no. 4, pp. 914–929, Jul. 2007.

[38] P. Y. Hao, "New support vector algorithms with parametric insensitive/margin model," *Neural Netw.*, vol. 23, no. 1, pp. 60–73, 2010.

[39] A. B. Ji, J. H. Pang, and H. J. Qiu, "Support vector machine for classification based on fuzzy training data," *Expert Syst. Appl.*, vol. 37, no. 4, pp. 3495–3498, Apr. 2010.

[40] S. Rajasegarar, C. Leckie, M. Palaniswami, and J. C. Bezdek, "Quarter sphere based distributed anomaly detection in wireless sensor networks," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2007, pp. 3864–3869.

[41] Y. Zhang, N. Meratnia, and P. Havinga, "An online outlier detection technique for wireless sensor networks using unsupervised quarter-sphere support vector machine," in *Proc. IEEE Int. Conf. Intell. Sensors, Sensor Netw., Inf. Process. (ISSNIP)*, Dec. 2008, pp. 151–156.

[42] L. Li, "An SVM-based abnormal events detection scheme in wireless sensor networks," *Comput. Appl. Softw.*, vol. 32, no. 2, pp. 272–277, Feb. 2015.

[43] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowl. Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[44] S. R. Gunn, "Support vector machines for classification and regression," *ISIS Tech. Rep.*, vol. 14, no. 1, pp. 5–16, 1998.

[45] S. M. Erfani, S. Rajasegarar, S. Karunasekera, and C. Leckie, "High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning," *Pattern Recognit.*, vol. 58, pp. 121–134, Oct. 2016.

[46] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," *ACM SIGMOD Rec.*, vol. 29, no. 2, pp. 427–438, 2000.

[47] F. Angiulli and C. Pizzuti, "Fast outlier detection in high dimensional spaces," in *Proc. Princ. Data Mining Knowl. Discovery (PKDD)*, Helsinki, Finland, 2002, pp. 15–26.

[48] B. Yu, M. Song, and L. Wang, "Local isolation coefficient-based outlier mining algorithm," in *Proc. Int. Conf. Inf. Technol. Comput. Sci.*, Jul. 2009, pp. 448–451.

[49] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, May 2000, pp. 93–104.

[50] T. Huang, Y. X. Zhu, Y. Wu, S. Bressan, and G. Dobbie, "Anomaly detection and identification scheme for VM live migration in cloud infrastructure," *Future Gener. Comput. Syst.*, vol. 56, pp. 736–745, Mar. 2016.

[51] T. Wang, J. Wei, W. B. Zhang, H. Zhong, and T. Huang, "Workload-aware anomaly detection for Web applications," *J. Syst. Softw.*, vol. 89, pp. 19–32, Mar. 2014.

[52] S. Kim, N. W. Cho, B. Kang, and S.-H. Kang, "Fast outlier detection for very large log data," *Expert Syst. Appl.*, vol. 38, no. 8, pp. 9587–9596, Aug. 2011.

[53] J. Tang, Z. Chen, A. W.-C. Fu, and D. W. Cheung, "Enhancing effectiveness of outlier detections for low density patterns," in *Proc. Pacific–Asia Conf. Adv. Knowl. Discovery Data Mining*, 2002, pp. 535–548.

[54] W. Jin, A. K. H. Tung, J. Han, and W. Wang, "Ranking outliers using symmetric neighborhood relationship," in *Proc. Pacific–Asia Conf. Adv. Knowl. Discovery Data Mining*, 2006, pp. 577–593.

[55] J. Nagi, K. S. Yap, S. K. Tiong, S. K. Ahmed, and A. M. Mohammad, "Detection of abnormalities and electricity theft using genetic support vector machines," in *Proc. IEEE Region 10 Conf. (Tencon)*, Nov. 2008, pp. 2176–2181.

[56] S. W. Fei and X. B. Zhang, "Fault diagnosis of power transformer based on support vector machine with genetic algorithm," *Expert Syst. Appl.*, vol. 36, no. 8, pp. 11352–11357, Oct. 2009.

[57] F. P. Yang, H. G. Wang, S. X. Dong, J. Y. Niu, and Y. H. Ding, "Two stage outliers detection algorithm based on clustering division," *Appl. Res. Comput.*, vol. 30, no. 7, pp. 1942–1945, Jul. 2013.

[58] G. C. Qian, R. Y. Jia, R. Zhang, and L. S. Li, "Outlier detection based on genetic algorithm for clustering," *Comput. Eng. Appl.*, vol. 44, no. 11, pp. 155–157, 2008.

[59] P. Gu, H. B. Liu, and Z. H. Luo, "Multi-clustering based outlier detect algorithm," *Appl. Res. Comput.*, vol. 30, no. 3, pp. 750–751, Mar. 2013.

[60] F. Jiang, Y. Sui, and C. Cao, "Some issues about outlier detection in rough set theory," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 4680–4687, Apr. 2009.

[61] X. Wang, X. Wang, Y. Ma, and D. Wilkes, "A fast MST-inspired kNN-based outlier detection method," *Inf. Syst.*, vol. 48, pp. 89–112, Mar. 2015.

[62] J. Nagi, K. S. Yap, S. K. Tiong, S. K. Ahmed, and M. Mohamad, "Nontechnical loss detection for metered customers in power utility using support vector machines," *IEEE Trans. Power Del.*, vol. 25, no. 2, pp. 1162–1171, Apr. 2010.

[63] S. S. S. R. Depuru, L. Wang, and V. Devabhaktuni, "Support vector machine based data classification for detection of electricity theft," in *Proc. Power Syst. Conf. Expo. (PSCE)*, Mar. 2011, pp. 1–8.

[64] F. Fabris, L. R. Margoto, and F. M. Varejao, "Novel approaches for detecting frauds in energy consumption," in *Proc. 3rd Int. Conf. Netw. Syst. Secur.*, Oct. 2009, pp. 546–551.

[65] E. W. S. dos Angelos, O. R. Saavedra, O. A. C. Cortés, and A. N. de Souza, "Detection and identification of abnormalities in customer consumptions in power distribution systems," *IEEE Trans. Power Del.*, vol. 26, no. 4, pp. 2436–2442, Oct. 2011.

[66] C. Cheng, H. Zhang, Z. Jing, M. Chen, L. Jiao, and L. Yang, "Study on the anti-electricity stealing based on outlier algorithm and the electricity information acquisition system," *Dianli Xitong Baohu Kongzhi/Power Syst. Protection Control*, vol. 43, no. 17, pp. 69–74, 2015.

[67] R. Vallakati, A. Mukherjee, and P. Ranganathan, "A density based clustering scheme for situational awareness in a smart-grid," in *Proc. IEEE Int. Conf. Electro/Inf. Technol. (EIT)*, May 2015, pp. 346–350.

[68] J. Nagi, K. S. Yap, S. K. Tiong, S. K. Ahmed, and F. Nagi, "Improving SVM-based nontechnical loss detection in power utility using the fuzzy inference system," *IEEE Trans. Power Del.*, vol. 26, no. 2, pp. 1284–1285, Apr. 2011.

[69] J. S. Lee and K. B. Lee, "An open-switch fault detection method and tolerance controls based on SVM in a grid-connected T-type rectifier with unity power factor," *IEEE Trans. Ind. Electron.*, vol. 61, no. 12, pp. 7092–7104, Dec. 2014.

[70] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The KDD process for extracting useful knowledge from volumes of data," *Commun. ACM*, vol. 39, pp. 27–34, Nov. 1996.

[71] J. Van den Broeck, S. A. Cunningham, R. Eeckels, and K. Herbst, "Data cleaning: Detecting, diagnosing, and editing data abnormalities," *PLoS Med.*, vol. 2, no. 10, pp. 966–970, Oct. 2005.

[72] D. Tanasa and B. Trousse, "Advanced data preprocessing for intersites Web usage mining," *IEEE Intell. Syst.*, vol. 19, no. 2, pp. 59–65, Mar. 2004.

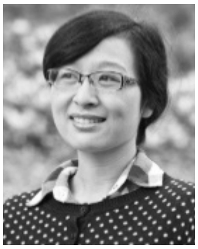
[73] S. Zhang, C. Zhang, and Q. Yang, "Data preparation for data mining," *Appl. Artif. Intell.*, vol. 17, nos. 5–6, pp. 375–381, 2003.

[74] X. Zhang, M. Wen, K. Lu, and J. Lei, "A privacy-aware data dissemination scheme for smart grid with abnormal data traceability," *Comput. Netw.*, vol. 117, pp. 32–41, Apr. 2017.

[75] Y. Yan, G. Sheng, Y. Chen, X. Jiang, Z. Guo, and X. Du, “An method for anomaly detection of state information of power equipment based on big data analysis,” *Zhongguo Dianji Gongcheng Xuebao/Proc. Chin. Soc. Elect. Eng.*, vol. 35, no. 1, pp. 52–59, 2015.
[76] M. Qi, Z. Fu, and F. Chen, “Outliers detection method of multiple measuring points of parameters in power plant units,” *Appl. Thermal Eng.*, vol. 85, pp. 297–303, Jun. 2015.
[77] J. Q. Gao, J. Y. Zhao, and X. Y. Fan, “Two improved data processing methods and their applications in SIS,” *Automat. Electr. Power Syst.*, vol. 33, no. 1, pp. 90–92, Jan. 2009.
[78] W. Sun and J. Hou, “A MPRM-based approach for fault diagnosis against outliers,” *Neurocomputing*, vol. 190, pp. 147–154, May 2016.
[79] K. Zhou and S. Yang, “Demand side management in China: The context of China’s power industry reform,” *Renew. Sustain. Energy Rev.*, vol. 47, pp. 954–965, Jul. 2015.
[80] S. Li and D. Zhang, “Developing smart and real-time demand response mechanism for residential energy consumers,” in *Proc. Power Syst. Conf. (PSC)*, Mar. 2014, pp. 1–5.
[81] B. Chakrabarti, D. Bullen, C. Edwards, and C. Callaghan, “Demand response in the New Zealand electricity market,” in *Proc. IEEE Transmiss. Distrib. Conf. Expo. (T&D)*, May 2012, pp. 1–7.
[82] Y. Ozturk, P. Jha, S. Kumar, and G. Lee, “A personalized home energy management system for residential demand response,” in *Proc. 4th Int. Conf. Power Eng., Energy Elect. Drives*, May 2013, pp. 1241–1246.
[83] K. Zhou, S. Yang, C. Shen, S. Ding, and C. Sun, “‘Energy conservation and emission reduction of China’s electric power industry,” *Renew. Sustain. Energy Rev.*, vol. 45, pp. 10–19, May 2015.
[84] L. Sun, K. Zhou, and S. Yang, “Regional difference of household electricity consumption: An empirical study of Jiangsu, China,” *J. Cleaner Prod.*, vol. 171, pp. 1415–1428, Jan. 2018.
[85] A. H. Nizar, J. H. Zhao, and Z. Y. Dong, “Customer information system data pre-processing with feature selection techniques for non-technical losses prediction in an electricity market,” in *Proc. Int. Conf. Power Syst. Technol.*, Oct. 2006, pp. 1–7.
[86] M. Farrokhifard, M. Hatami, and M. Parniani, “Novel approaches for online modal estimation of power systems using PMUs data contaminated with outliers,” *Electr. Power Syst. Res.*, vol. 124, pp. 74–84, Jul. 2015.
[87] S. Jamali, A. Bahmanyar, and E. Bompard, “Fault location method for distribution networks using smart meters,” *Measurement*, vol. 102, pp. 150–157, May 2017.
[88] N. H. Yu, L. F. Li, L. Wang, H. G. Yang, and D. Tan, “Abnormal detection for harmonic currents based on cloud model,” *Proc. CSEE*, vol. 34, no. 25, pp. 4395–4401, Sep. 2014.
[89] Y. Wang, Y. M. Zhang, P. H. Qian, and L. Shen, “On relationship between electricity grid system load and meteorological conditions in Suzhou and forecasting system design,” *J. Anhui Agricult. Sci.*, vol. 42, no. 13, pp. 3959–3961, 2014.
[90] A. Kenward and U. Raja. (2014). *Blackout: Extreme Weather, Climate Change and Power Outages*. [Online]. Available: http://www.ourenergypolicy.org/wp-content/uploads/2014/04/climate-central.pdf
[91] L. Wigle. (2014). *How Big Data Will Make us More Energy Efficient*. [Online]. Available: https://www.weforum.org/agenda/2014/05/big-data-will-make-us-energy-efficient/?utm_content=buffera432a&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer2014
[92] D. M. Ward, “The effect of weather on grid systems and the reliability of electricity supply,” *Climatic Change*, vol. 121, no. 1, pp. 103–113, Nov. 2013.
[93] U. G. Knight, “Power systems in emergencies: From contingency planning to crisis management,” *Int. J. Elect. Eng. Educ.*, vol. 38, p. 382, Oct. 2001.
[94] K. Zhou, C. Fu, and S. Yang, “Big data driven smart energy management: From big data to big insights,” *Renew. Sustain. Energy Rev.*, vol. 56, pp. 215–225, Apr. 2016.
[95] B. Lu and W. Song, “Research on heterogeneous data integration for smart grid,” in *Proc. 3rd IEEE Int. Conf. Comput. Sci. Inf. Technol.*, Jul. 2010, pp. 52–56.
[96] T. Liu *et al.*, “Abnormal traffic-indexed state estimation: A cyber–physical fusion approach for Smart Grid attack detection,” *Future Gener. Comput. Syst.*, vol. 49, pp. 94–103, Aug. 2014.
[97] V. P. Janeja and V. Atluri, “Spatial outlier detection in heterogeneous neighborhoods,” *Intell. Data Anal.*, vol. 13, no. 1, pp. 85–107, 2009.
[98] M. Solaimani, M. Iftekhar, L. Khan, and B. Thuraisingham, “Statistical technique for online anomaly detection using Spark over heterogeneous data from multi-source VMware performance data,” in *Proc. IEEE Int. Conf. Big Data*, Oct. 2014, pp. 1086–1094.
[99] J. I. Guerrero, A. García, E. Personal, J. Luque, and C. León, “Heterogeneous data source integration for smart grid ecosystems based on metadata mining,” *Expert Syst. Appl.*, vol. 79, pp. 254–268, Aug. 2017.
[100] J. Soares, M. A. F. Ghazvini, M. Silva, and Z. Vale, “Multi-dimensional signaling method for population-based metaheuristics: Solving the large-scale scheduling problem in smart grids,” *Swarm Evol. Comput.*, vol. 29, pp. 13–32, Aug. 2016.
[101] P. D. Diamantoulakis, V. M. Kapinas, and G. K. Karagiannidis, “Big data analytics for dynamic energy management in smart grids,” *Big Data Res.*, vol. 2, no. 3, pp. 94–101, Sep. 2015.
[102] B. U. Kim, C. Lynn, N. Kunst, S. Vohnout, and K. Goebel, “Fault classification with Gauss–Newton optimization and real-time simulation,” in *Proc. Aerosp. Conf.*, Mar. 2011, pp. 1–9.
[103] B. U. Kim and S. Hariri, “Anomaly-based fault detection system in distributed System,” in *Proc. 5th ACIS Int. Conf. Softw. Eng. Res., Manage., Appl.*, Aug. 2007, pp. 782–789.
[104] C. C. O. Ramos, A. N. de Sousa, J. P. Papa, and A. X. Falcao, “A new approach for nontechnical losses detection based on optimum-path forest,” *IEEE Trans. Power Syst.*, vol. 26, no. 1, pp. 181–189, Feb. 2011.
[105] Z. Xue, Y. Shang, and A. Feng, “Semi-supervised outlier detection based on fuzzy rough C-means clustering,” *Math. Comput. Simul.*, vol. 80, no. 9, pp. 1911–1921, May 2010.
[106] A. Bulut *et al.*, “Real-time nondestructive structural health monitoring using support vector machines and wavelets,” *Proc. SPIE*, vol. 5770, pp. 180–189, May 2005.
[107] D. J. Hill and B. S. Minsker, “Anomaly detection in streaming environmental sensor data: A data-driven modeling approach,” *Environ. Modell. Softw.*, vol. 25, no. 9, pp. 1014–1022, Sep. 2010.
[108] X. Sun, Y. Chen, J. Liu, and S. Huang, “A co-simulation platform for smart grid considering interaction between information and power systems,” in *Proc. Innov. Smart Grid Technol. Conf. (ISGT)*, Feb. 2014, pp. 1–6.
[109] A. Sancho-Asensio *et al.*, “Improving data partition schemes in Smart Grids via clustering data streams,” *Expert Syst. Appl.*, vol. 41, no. 13, pp. 5832–5842, Oct. 2014.
[110] S. K. Card, J. D. Mackinlay, and B. Shneiderman, *Readings in Information Visualization: Using Vision to Think*. San Mateo, CA, USA: Morgan Kaufmann, 1999.
[111] P. M. Mahadev and R. D. Christie, “Case study: Visualization of an electric power transmission system,” in *Proc. IEEE Conf. Vis.*, Oct. 1994, pp. 379–381.
[112] Y. Sun and T. J. Overbye, “Visualizations for power system contingency analysis data,” *IEEE Trans. Power Syst.*, vol. 19, no. 4, pp. 1859–1866, Nov. 2004.
[113] T. J. Overbye, “Power system visualization,” *Dianli Xitong Zidonghua/Automat. Electr. Power Syst.*, vol. 29, no. 16, pp. 60–65, 2005.
[114] J. D. Weber and T. J. Overbye, “Voltage contours for power system visualization,” *IEEE Trans. Power Syst.*, vol. 15, no. 1, pp. 404–409, Feb. 2000.
[115] C. J. Guo *et al.*, “Application of scientific visualization to power systems,” *Water Resour. Power*, vol. 29, no. 2, pp. 46–149, Feb. 2011.

LI SUN received the B.S. degree from the School of Management, Hefei University of Technology, Hefei, China, in 2016, where she is currently pursuing the Ph.D. degree with the School of Management. Her current research interests include demand response, clustering method, data mining, and smart energy management.

KAILE ZHOU received the B.S. and Ph.D. degrees from the School of Management, Hefei University of Technology, Hefei, China, in 2010 and 2014, respectively. From 2013 to 2014, he was a Visiting Scholar with the Eller College of Management, The University of Arizona, Tucson, AZ, USA. He is currently an Associate Professor with the School of Management, Hefei University of Technology, and a Post-Doctoral Research Fellow with the City University of Hong Kong, Hong Kong. He has authored in the *Applied Energy*, the *Renewable and Sustainable Energy Reviews*, and the *Journal of Cleaner Production* among others. His research interests include smart energy systems, energy informatics, and big data analytics.



XIAOLING ZHANG received the B.S. degree from Shandong University and the Ph.D. degree from The Hong Kong Polytechnic University. She was with The University of Hong Kong and The Hong Kong Polytechnic University. She is currently an Associate Professor with the Department of Public Policy, City University of Hong Kong. She has authored over 110 referred journal papers. Her research interests include sustainability science, energy consumption behavior, energy policy, environmental studies, and facility management.

SHANLIN YANG is currently a Distinguished Professor with the School of Management, Hefei University of Technology, Hefei, China. He has authored over 300 referred journal papers and over 200 conference papers. His research interests include engineering management, smart energy management, and strategic management. He is a member of the Chinese Academy of Engineering. He is a fellow of the Asian–Pacific Industrial Engineering and Management Society. He is also the Vice Chairman of the China Branch of the Association of Information Systems.