Zhang et al. Energy Informatics (2018) 1:8
https://doi.org/10.1186/s42162-018-0007-5

Energy Informatics

REVIEW Open Access

# Big data analytics in smart grids: a review

Yang Zhang, Tao Huang* and Ettore Francesco Bompard

* Correspondence: tao.huang@polito.it
Department of Energy, Polytechnic University of Turin, Corso Duca degli Abruzzi, 24, 10129 Torino, Italy

## Abstract

Data analytics are now playing a more important role in the modern industrial systems. Driven by the development of information and communication technology, an information layer is now added to the conventional electricity transmission and distribution network for data collection, storage and analysis with the help of wide installation of smart meters and sensors. This paper introduces the big data analytics and corresponding applications in smart grids. The characterizations of big data, smart grids as well as huge amount of data collection are firstly discussed as a prelude to illustrating the motivation and potential advantages of implementing advanced data analytics in smart grids. Basic concepts and the procedures of the typical data analytics for general problems are also discussed. The advanced applications of different data analytics in smart grids are addressed as the main part of this paper. By dealing with huge amount of data from electricity network, meteorological information system, geographical information system etc., many benefits can be brought to the existing power system and improve the customer service as well as the social welfare in the era of big data. However, to advance the applications of the big data analytics in real smart grids, many issues such as techniques, awareness, synergies, etc., have to be overcome.

## Introduction

With the fast development of digital technology and cloud computing, more and more data are produced through digital equipment and sensors, such as smart phones, computers, advanced measuring infrastructures, etc., as well as through human activities and communications. For instance, the size of data on the internet is now measured in exabytes ($10^{18}$) and zettabytes ($10^{21}$) (Emani et al., 2015). Rational, effective and efficient analysis of these data brings huge value and benefit to our daily life and company activities. However, the collected data are mounting at an exponential growth, and the structure of them is also becoming much more complicated. The processing and analysis method of these large volume data is a new challenge but opportunity at the beginning of this century with the concept of "big data" (Lv et al., 2017a; Günther et al., 2017).

Although big data is a newly-appeared term, the concept of discovering valuable information from massive collected data in commercial operation as aiding knowledge for business decision has already been proposed in 1989 by Howard Dresner as "business intelligence" (BI) (Yu, 2002). The trend of internet revolution and ubiquitous information acquisition devices successfully reduce the cost of data collection, while the huge amount and complex structure challenge the capability of traditional data analytics techniques.

In power grid, the traditional fossil fuels are facing the problem of depletion and the de-carbonization demands the power system to reduce the carbon emission. Smart grid and super grid are effective solutions to accelerate the pace for electrification of human society with high penetration of renewable energy sources (Ak et al., 2016). Although the rising awareness of sustainable development have become the impetus to the utilization of renewable energy sources, the intermittent characteristics of wind and photovoltaic energies bring huge challenges to the safe and stable operation in a low inertia power system (Wenbin & Peng, 2017; Ye et al., 2016). The data analytics based renewable energy forecasting methods are a hot research topic for a better regulation and dispatch planning in such cases. Traditional electricity meters in distribution systems only produce a small amount of data which can be manually collected and analyzed for billing purpose. While the huge volume of data collected from two-way communication smart grids at different time resolutions in nowadays need advanced data analytics to extract valuable information not only for billing information but also the status of the electricity network. For example, the high-resolution user consumption data can also be used for customer behavior analysis, demand forecasting and energy generation optimization. Predictive maintenance and fault detection based on the data analytics with advanced metering infrastructure are more crucial to the security of power system (Chunming et al., 2017).

Thus, the great progress of information and communication technology (ICT) provides a new vision for engineers to perceive and control the traditional electrical system and makes it smart. An embedded information layer into the energy network produces huge volume of data, including measurements and control instructions in the grid for collection, transmission, storage and analysis in a fast and comprehensive way. It also brings a lot of opportunities and challenges to the data analysis platform. This paper is to discuss the concepts of data analysis and their applications in smart grids. The intent of this paper is three-fold. First the potential data collected with advanced metering infrastructure in smart grid are discussed. Next, the paper briefly reviews the concepts of data analytics and the popular techniques. Finally, the paper illustrates the detailed applications of data analytics in smart grid.

## Big data in smart grid

### Concept of big data

The definition of big data is not very clear and uniform at present. But there is a consensus among different descriptions: this is an emerging technical problem brought by a dataset of large volume, various categories and complicated structures which needs novel framework and techniques to excavate useful information effectively. Therefore, the definition of big data depends on the ability of data mining algorithms and the corresponding hardware equipment to deal with large volume datasets (Zikopoulos & Eaton, 2011). It is a relative concept instead of an absolute definition. The big data can be understood as amount of data beyond technology's capability to store, manage and process efficiently in (Kaisler et al., 2012) as the data size increasing along with the evolvement of ICT technologies.

### Concept of smart grids

Smart grid is the power system embedded with an information layer that allows for two-way communication between the central controllers and local actuators as well as logistic units to respond digitally to urgent situations of physical elements or quickly

changing of electric demand. The E.U. defined the smart grid as electricity networks that can intelligently integrate the actions of all users connected to it – generators, consumers and those that do both – in order to efficiently deliver sustainable, economic and secure electricity supplies (SmartGrids European Tech, 2010). The U.S. defined the smart grid of future in a similar way that incorporates the digital technology to improve reliability, security and efficiency of the electric system through information exchange, distributed generation and storage resources for a fully automated power delivery network (Zhen Zhang. Smart Grid in America and Europe, 2011).

Compared with traditional power systems, the widespread application of distributed generators under the call of green energy resources is shaking the hegemony position of large-scale centralized power plants, which makes the conventional centralized control strategy less effective due to the unidirectional power flow. Connection of small-scale power generations (typically in the range of 3 kW to 10 kW) to the public distribution grid requires two-directional operation and control of distribution grids. Faced with the challenges of more complicated control and protection strategies, the conventional electro-mechanical electric grid is supposed to be enhanced with the help of innovations in the digital information and telecommunications network to overcome the cost from power outages and power quality disturbances as billions of dollars annually (Executive Office of the President, 2013).

Normally, the smart grid can be assessed with a Smart Grid Architecture Model (SGAM), which is a 3-dimensional framework that merges domains, zones and layers together. The conventional structure of power system can be found in the domains as generation, transmission, distribution, DER (Distributed Energy Resources) and customer premises. The zones which present the layout of power system management are composed of market, enterprise, operation, station, field and process. On top of the first two dimensions, the layout of interoperability layers includes the component, communication, information, function and business layers. SGAM as an architectural overview can be used to find the limitations and commonalities of existing smart grid standards (CEN-CENELEC-ETSI Smart Grid Working Group Reference Architecture, 2012).

### Big data characteristics in smart grid

The characteristics of big data in smart grid is also in accordance with the universal 5 V big data model in many researches (Zhu et al., 2015) as below:

(i) Volume – refers to the vast amount of data generated, which makes data sets too large to store and analyze using traditional database technology. The possible solution to this problem is the distributed systems to store data in different locations, connect them by networks and bring them together by software. In smart grid the widespread application of smart meter and advanced sensor technology provide huge amount of data.

(ii) Velocity – refers to the speed at which new data is generated and the speed at which data moves around. The requirements for real-time exchange of data is increasing. With a sampling rate of 4 times per hour, 1 million smart meters installed in the smart grid would result in 35.04 billion records, equivalent to 2920 Tb data in quantification (SAGIROGLU et al., 2016). The following Table 1 indicates the

**Table 1** Quantification of collected data in different sampling rates (Big Data analytics and energy consumption, 2016)

| Collection Frequency | 1/day | 1/h | 1/30 min | 1/15 min |
|---|---|---|---|---|
| Records Collected | 365 million | 8.75 billion | 17.52 billion | 35.04 billion |
| Volume of Data | 1.82 TB | 730 TB | 1460 TB | 2920 TB |

amount of records from smart meters in a year under various collection frequency with the assumption of 1 million devices and a 5 KB record per collection.

(iii) Variety – refers to the types of data we can now use. In the past, we focus on structured data that neatly fits into tables or rational databases such as financial or meteorological data. With big data technology, we need to handle different types of unstructured data including messages, social media conversations, digital images, sensor data, video or voice recordings, and bring them together with more traditional, structured data. According to the extensive data sources in smart grid as shown in Fig. 1, the formats and dimensions of data are diverse in structure.

(iv) Veracity – refers to the messiness or trustworthiness of the data. The quality and accuracy are less trustworthy with such large amount of big data, which challenge the outcome data analysis. Errors of measurements in smart grid may exist due to the imperfections in devices or mistakes in data transmission. The secure and efficient power system operation relays on the data assessment and state estimation.

(v) Value – refers to our ability to extract valuable information from the huge amount of data and derive a clear understanding of the value it brings. The larger the amount of data is, the lower the density of valuable information will be. With the improvement of intelligent devices adopted in smart grid, more and more value of big data analytics is revealed according to the various applications.

### Data sources in smart grids

As an intelligent system of both energy and information, smart grid is the abundant source of information, which covers the data from process of electricity generation, transmission, distribution and consumption. These data include the electrical information from distribution stations, distribution switch stations, electricity meters, and non-electrical information like marketing, meteorological as well as reginal economic data as shown in Fig. 1 (Keyan et al., 2015). Collection and analysis of them provide essential help in scheduling of power plants, operation of subsystems, maintenance for vital power equipment and business behavior in marketing.

The data sources mentioned above can be sorted into three categories: measurement data, business data and external data (Teng et al., 2014). Most of the operation parameters in power system are measured through installed sensors and smart meters, indicating the system's current and historical status (SAGIROGLU et al., 2016) (Jiye et al., 2015). The weather conditions and social events like festivals are the external data that cannot be measured from smart meters but have an impact on the operation and planning in power system. The business data mainly includes the marketing strategies and rivals' behaviors.

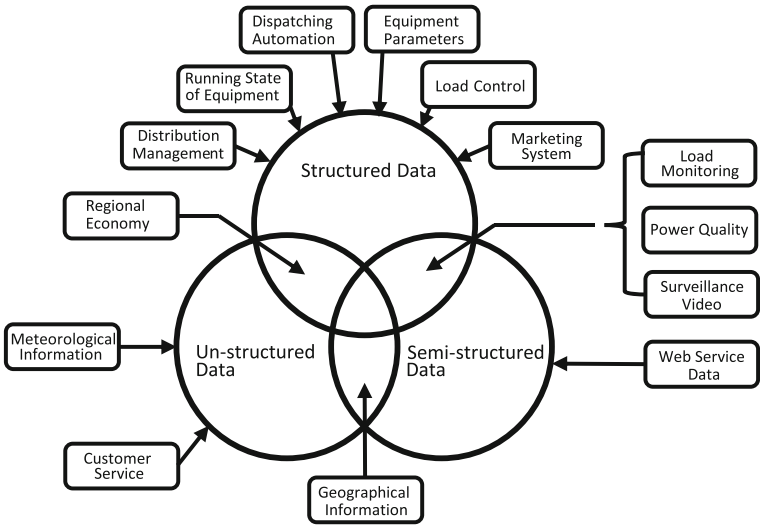


**Fig. 1** Data Sources of the Grids

### Data collection techniques in smart grid

In smart grid, the data are collected and transmitted with help of smart meters which provide energy related information to both the utility company (or DSO) and customers. For the energy consumption of residential customers, the number of smart meter readings for a large utility company is expected to rise from 24 million a year to 220 million per day (SAGIROGLU et al., 2016). As an emerging component in electricity market and smart grid, electric vehicles (EVs) and plug-in hybrid EVs (PHEVs) have seen a growing popularity with the movement of electrification in transportation sector and progress of artificial intelligence. To control the normal operation status of the distribution system, DSO traditionally relies on the measurements in the primary substation, at the beginning of each MV feeder, where the protection systems are normally installed. The current magnitude information is also needed for the automatic on-load tap changer in HV/MV transformers for voltage regulation. The measurements of a typical smart meter include the node voltage, feeder current, power factor, active and reactive power, energy over a period, total harmonic distortion as well as load demand, etc. The intelligent devices for data collection in smart grid are listed as Table 2.

### Data communication techniques in smart grid

The communication infrastructure of the smart grid is composed of three types of networks: home area network (HAN), neighborhood area network (NAN) and wide area network (WAN) as shown in Fig. 2 (Baimel et al., 2016). The functions and characteristics of the above communication infrastructures are summarized in Table 3.

Basic types of communication technologies for smart meters include wired and wireless infrastructures. The wireless communication technology allows the data center to gather measurement information from smart meters with low costs and simple connections while it may face the electromagnetic problem. Power line communication (PLC) is a wired communication technology by add a modulated carrier signal to the power cables and already successfully implemented in power system. The existing communication

**Table 2** Intelligent data collection devices in smart grid

| Intelligent device | Technology | Application |
|---|---|---|
| Advanced metering infrastructure (AMI) | Integration of smart meters, data management systems and communication networks to provide bidirectional communication between customers and utilities. | Remote meter configuration, dynamic tariffs, power quality monitoring and local control |
| Phasor measurement unit (PMU) | Real-time measurements (30 to 60 samples/second) of multiple remote points with a common time source for synchronization | Electrical waves measurement of power grid |
| Wide area monitoring system (WAMS) | An application server to deal with the incoming information from PMUs | Dynamic stability of the grid |
| Remote terminal unit (RTU) | A microprocessor-controlled device that transmitting telemetry data | Information collection of system operation status |
| Supervisory control and data acquisition (SCADA) | Both manual and automatic | System monitoring, event processing and alarm |
| Intelligent electronic device (IED) | Monitoring and recording status changes in the substation and outgoing feeders | Combination of different relay protection functions with measurement, recording and monitoring |

technology include ZigBee, WALN, cellular communication, WiMAX, PLC, etc. (Baimel et al., 2016).

As one of the first countries for smart metering infrastructure development, Italy has deployed smart meter to nearly all the customers with the PLC technology to transfer smart meter data to the nearest data concentrator located in the MV/LV substation. Then these data are sent to the DSO's data centers for recording and data analysis.

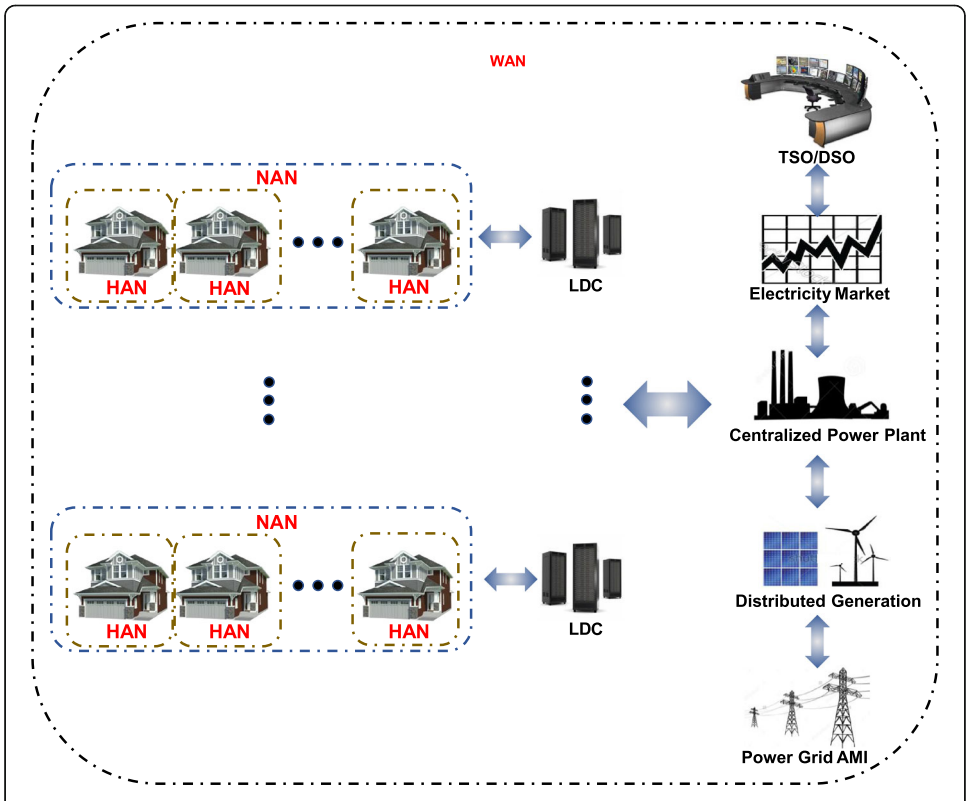


**Fig. 2** Smart grid communication infrastructure

**Table 3** Summary of communication infrastructure in smart grid

| Type of network | Function | Characteristic |
|---|---|---|
| HAN | Enabling the communication among smart home or office devices and smart meters for local energy management | Deployed at house or small office with a relatively low transmission data rate (less than 1 Kbps) |
| NAN | Consisting of several HANs for energy consumption data aggregation and storage at load data center (LDC) | Deployed within area of hundreds of meters with up to 2Kbps |
| WAN | Enabling the communication of all smart grid's components | Deployed within tens of kilometers with high data transmission capability up to few Gbps |

There are around 30 million meters and 400,000 secondary substation concentrators installed (Bahmanyar et al., 2016).

## Data analysis techniques

The most important stage of the big data processing system is data analysis, which is the basis for discovering valuable information and supporting the decision-making (Fan et al., 2018; Cheng et al., 2018). There are several similar concepts relevant to data analysis listed in Table 4.

From a general point of view, the data analytics or data mining is the computational process to reveal the potential relations between variables with the techniques including database, statistics, pattern recognition, machine learning, etc. However, due to the diverse sources, the collected data sets may have different levels of quality in terms of noise, redundancy and consistency.

### Data preprocessing

The data pre-processing techniques are necessary to improve data quality as shown in Fig. 3.

Data integration techniques aim to aggregate data collected from disparate sources in an effective way with a unified view (Roya et al., 2018). For example, when combining the datasets of weather condition records and power system interruption events, the attribute of "date time" would appear twice. But apparently only one attribute of "date time" is needed for the following data analytics process. The same attributes with different name as well as the different attributes with the same name is to be identified in this process (Lim et al., 1996). Normally, the correlation analysis is used in the redundancy

**Table 4** Concepts related to data analysis

| Concept | description |
|---|---|
| Statistics | The study of data collection, analysis and interpretation with mathematics methods which may discover potential relations based on some hypothesis |
| Machine learning | A kind of technique for understanding the law in the data as well as extracting useful information with the help of computers automatically instead of humanity |
| Data mining | Computing data for discovering valuable information in large data sets with knowledge of statistics, machine learning and database system. |
| Pattern recognition | A branch of machine learning that focuses on the regularities in data |
| Deep learning | A branch of machine learning based on complex structure of neural networks |
| Artificial intelligence | The study of intelligent systems and agents with the ability of learning from circumstances and solving problems |

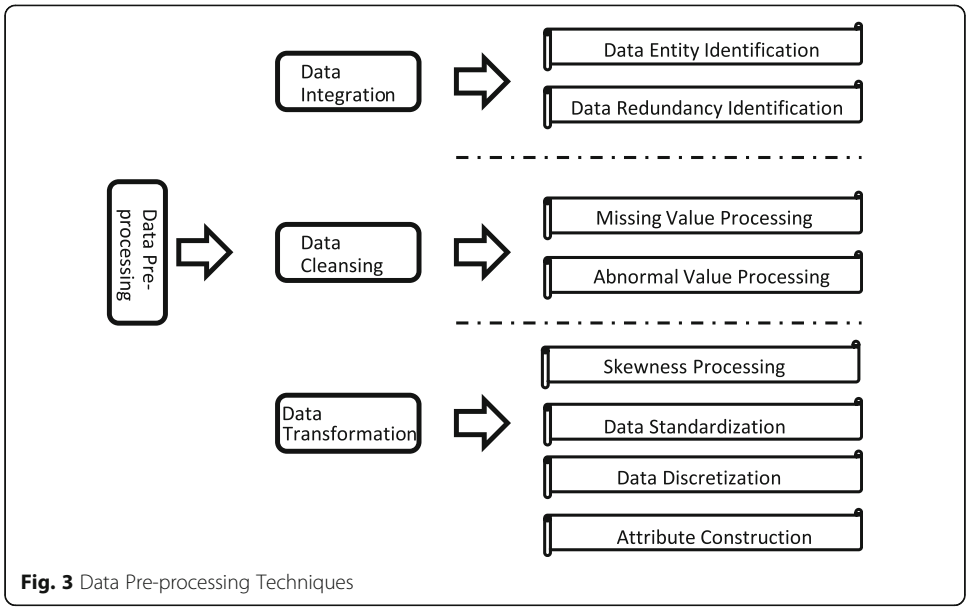


**Fig. 3** Data Pre-processing Techniques

identification to abandon the highly correlated attributes and reduce the size of datasets. In most cases, the datasets would contain some missing values which influence the results of data analytics. Deletion or interpolation are the frequent techniques to solve such kind of problems. As to the abnormal values, the first step is to check whether this is rational based on the professional knowledge for the application. If it is caused by an error in sensors or data processing platform, we can treat it as a missing value or try to find the real value, otherwise it is supposed to be kept in the dataset as a "black swan". The logarithm is an effective way to "correct" the distribution shape of data with severe skewness, because some data analytics algorithms are sensitive to imbalanced data. New attributes such as the temperature difference can be calculated in the pre-processing step if there is only maximum and minimum value of temperature in the initial dataset. The new constructed attributes are usually helpful to improve the accuracy of data analytics results.

### Data analytics techniques

The most frequently used data mining or machine learning algorithms are usually categorized as supervised or unsupervised learning depending on whether there is a label attached to each item in datasets as shown in Table 5. For the supervised learning algorithms, the data analytics model can be trained based on the given data to discover the relation between data attributes and the corresponding categories or values. While for those without labels, the data analytics model is usually designed to recognize the possible groups among all the items (Di Zhua & Zhang, 2018).

### Procedures of data Mining in Smart Grids

As shown in Fig. 4, the main procedure of data analytics in smart grid is to extract valuable information from historical data for guiding the operation and maintenance with the comparison to real-time data (Siryani et al., 2017). The huge amount of data collected from smart meters and sensors are arranged and stored with data management techniques. After preparation, the mathematical model can be established

**Table 5** Data Analytics Algorithms

| Category | Algorithm | description |
|---|---|---|
| Supervised Learning | Decision tree | A non-parametric method with a tree-like method whose leaves represent class labels and branches represent conjunctions of features |
| | Naive Bayes | A probabilistic method based on Bayes theorem with the assumption of independence between every pair of features |
| | Support vector machine classifier | An algorithm to find a separating hyperplane between the two classes by mapping the labelled data to a high-dimensional feature space |
| | K Nearest Neighbor | A non-parametric method based on the minimum dissimilarity between new items and the labelled items in different classes |
| | Random Forest | An algorithm consisting of a collection of simple tree predictors independently for the estimation of the final outcome |
| Unsupervised Learning | K-means | An unsupervised learning method with a given number of clusters to sort the data based on the average value of data in each group as the centroid |
| | K-medoids | An unsupervised learning method similar to k-means by assigning the centroid of each group with an existing data point instead of the average value |
| | Hierarchical Clustering | An alternative approach which aims to build a hierarchy of clusters in a dendrogram without a given number of clusters |
| | DBSCAN | A density-based clustering algorithm to identify clusters with specific shape in distribution |
| | Expectation-Maximization | An iterative way to approximate the maximum likelihood estimates for model parameters |
| Correlation | FP-Growth Algorithm | An efficient method for mining the complete set of frequent patterns with a special data structure named frequent-pattern tree with all the association information reserved |
| | Apriori Algorithm | A classical data analytics algorithm to discover the potential association rules among frequent items |
| Dimensionality reduction | Principal Component Analysis | An orthogonal transformation of data with a new coordinate system with the greatest variance projected to the first coordinate |
| | Self-organizing Map | A type of artificial neural network for a low-dimensional representation of the training data space |
| | Random Matrix | An algorithm which reveal potential regulations with high order matrices for massive data by eigenvalue analysis |

through data mining techniques based on the clean data. With the input of real-time measurements, the state status can be evaluated in the derived model, which provides the possible schemes to guide practical actions and solve potential problems.

## Big data analytics in smart grid

### Fault detection

The carbon emission reduction and sustainability of environment are the driving force and construction purpose of smart grid, which is designed in a decentralized structure. The employment of distributed generator units in modern power distribution system now provides an effective means for the utilization of widespread renewable energy such as

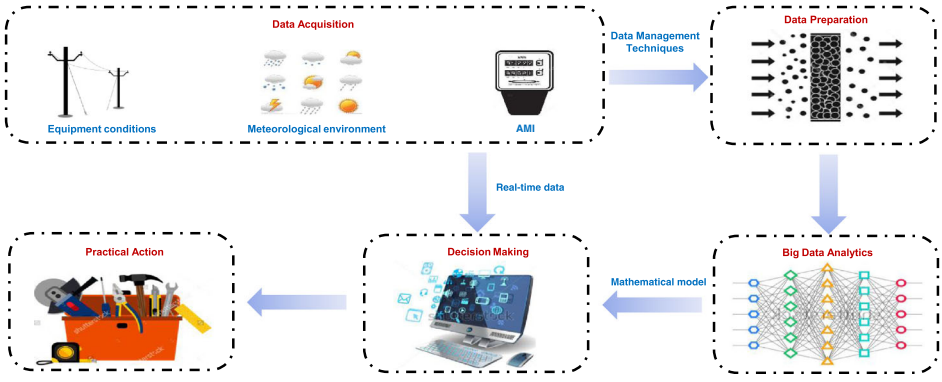


**Fig. 4** Example of big data analytic procedures in smart grid

wind and solar energy. These emerging microgrids are vital for the expectation of a low-carbon society. Moreover, the close distance between the generator and loads in microgrid improves the reliability of power delivery and reduces the power transmission loss. The ability to operate in an island mode also protects the load from damages caused by power system including voltage fluctuation, frequency deviation, etc. (Mishra et al., 2016).

However, the intermittent characteristic of renewable energy increases the uncertainty in power grid, whose typical solution is to use inverter interfaced distributed generators (IIDGs) for a better power quality. In contrast with the traditional bulk generators like large-volume thermal, nuclear or hydro generators, the much lower inertia of IIDGs is a severe potential threat when the faults in microgrids cannot be detected and cleared in a short time due to the limited current carrying capacity. Most of the traditional techniques relying on the detection of overcurrent and negative sequence current origin from the large-scale centralized power system and seem less effective in microgrids. A statistical classifier-based protection scheme using local current measurements is proposed by applying the wavelet transform and the decision tree (DT) model in (Mishra et al., 2016). The wavelet transform can decompose the signal in time-frequency domain with the time localization reserved. Energy, Shannon entropy and standard deviation of the wavelet coefficients which contain the information during transient events are calculated. Finally, 15 statistical features extracted from the current data for one cycle by sequence analyzer and wavelet transformation are fed into the DT models for fault detection and classification. A differential protection scheme for microgrid is proposed in (Kar et al., 2017) with the most sensitive features at both ends of the respective feeder processed by the discrete Fourier transform. These differential features are then utilized in the decision tree-based data-mining model for determining the final relaying decision.

For a grid-connected microgrid, the severe weather conditions or grid blackouts may trigger an unintentional islanding accident, which threats the safety operation and causes technical issues. Artificial neural networks (ANNs) are trained in (Hashemi et al., 2017) with features extracted from the differential transient of the rate of change of frequency (ROCOF) signal in order to identify islanding accidents. A support vector machine (SVM) classifier is established in (Alam et al., 2017) with multiple features extracted from system variables as an islanding detection approach. The feature extraction process is implemented with a sliding window whose width is optimized for the highest detection rate.

As a real-time social sensor for the smart grid, social media like Twitter or Facebook could contain potential information indicating the occurrence and location of power outages (Bauman et al., 2017). A probabilistic framework is devised in (Sun et al., 2016) for detecting a targeted event from the fragmented and noisy tweets. The method shows a good performance in locating accrual outage areas in experiment, which could be integrated to a social data-driven outage management.

#### Predictive maintenance/condition based maintenance

Distribution automation (DA) is a concept of smart grid which focuses on the operation and system reliability at the distribution level. A successful DA has the capability to localize and isolate the faults in distribution system with a reduced restoration time and improved customer satisfaction. Under the concept of DA, increasing volume of operational data have been collected from supervisory control and data acquisition (SCADA) or advanced metering infrastructure (AMI) for state monitoring and fault diagnosis.

Reference (Wang et al., 2017a) proposes an analyzing scheme for preventative measures to avoid or minimize the outages with the data related to pole mounted auto-recloser (PMAR). PMAR is a kind of protection intelligent electronic device installed on the overhead lines of a distribution network which attempts several recloses after an interruption happened in the downstream of the feeder.

Thanks to the development of ICT technology in power systems, a huge volume of data can be collected via AMI and communication infrastructures. Power system operating data, weather information and log data of relay protection devices are processed as the input of a one class classification system, which is a data-driven model of fault phenomena based on a hybridization of evolutionary learning and clustering techniques in (De Santis et al., 2015; De Santis et al., 2017). This fault recognition system is validated in the medium voltage power grid in Rome. The traditional statistical methods such as linear discriminant analysis (LDA) and logistic regression are discussed for mining the relation between power system faults and the features extracted from raw data (Cai & Chow, 2009).

As a potential threat to the security of transmission systems, the galloping of power lines can cause structural and electrical failures. After analyzing the impact factors of galloping, a data-driven model based on SVM and AdaBoost bi-level classifiers is proposed in (Wang et al., 2016a) for early warning. The extreme learning machine (ELM) algorithm is applied in an intelligent early-warning system for reliable online detection of risky events in power system in (Zhang et al., 2017). Since the weights in ELM training are randomly chosen and then determined through matrix computation without iterative parameter adjustment, the learning speed is much faster than conventional algorithms, which is an ideal solution in “big data” cases. The optimal balance between earning accuracy and warning earliness of the data-driven framework is also discussed. Reference (Cui et al., 2017) provides a method to extract electrical features from high-impedance fault current and voltage signals and build an effective feature set (EFS) via a ranking algorithm. Therefore, only a small number of signal channels are required to build a statistical classifier for fault detection. Reference (Jiang et al., 2016) also provides an effective method to reduce the huge volume of PMU data while retaining the critical information for fault detection in power system.

### Transient stability analysis

Transient stability is a critical issue closely related to the safely operation of power system. However, the increasing demand for electricity, growing penetration of renewable energy sources and deregulated market force power grid to operate near their secure operating limits (Liu et al., 2014). Facing with the challenges from a more complex system, transient stability analysis (TSA) for the study of dynamic behavior taking the electromechanical and electromagnetic process in power system taken into consideration is becoming a hot research topic. The transient process and new operating conditions need be calculated with the TSA technique after a severe interruption in power grid for a comprehensive protection scheme. Traditional TSA based on the time-domain simulation is not able to provide universal results due to so many uncertainties.

Under the concept of smart grid, a large amount of data collected via AMI are involved in the state assessment of power systems to support the energy management, system operation and decision making. Therefore, efficient summarization techniques are required for extracting useful patterns and discovering valuable information from redundant measurements in power system. A DT-based framework is proposed in (Liu et al., 2014) (Vittal, 2013) for the dynamic security assessment (DSA) in power system with high penetration of DGs. Two contingency-oriented DTs are trained based on the databases generated from real-time simulations. One of the well-trained DT is fed with real-time wide-area measurements to identify potential security issues, and the other DT provides the online corresponding preventive control strategies to deal with the problems. In (He et al., 2016) the dominant instability generation group (DIGG) in power system is identified without time domain simulation since the features adopted for TSA are extracted from steady-state variables. Reference (Parate et al., 2016) proposed an approach to classify the collected data from smart grid into two classes called vulnerable and non-vulnerable data sets with the data analytics such as multichannel singular spectrum analysis (MSSA), principal component analysis (PCA) and SVM. A framework for online contingency screening is presented in (Dimitrovska et al., 2017) with respect to first swing transient stability. The large spectrum of pre-fault operating state variables and critical clearing times of several contingencies are collected to compose a dataset for pattern recognition methods. The metric which can be used for operating condition evaluation is developed through PCA.

In addition to the renewable energy micro-sources distributed in smart grid (SG), the grid-connected high capacity wind farms are also widely accepted and applied for an effective utilization of pollution free and abundant nature resources. The improvement of technologies for large wind turbine generators and high capacity power converters accelerate the amount of wind energy integration into power system. To address the potential deterioration and stability problem caused by the large integration of wind energy to power grid, reference (Andalib-Bin-Karim et al., 2017) proposed data-driven analytics to determine the Q-V characteristic curve at the point of interconnection of the wind farm with valuable information for voltage stability extracted. Without prior knowledge of the system configuration and parameters, different curve-fitting techniques are adopted in a real case study in Canada.

Power swing is the oscillation of power flow on transmission lines when the angles of rotors of synchronous machines are advancing or retracting to each other which may cause a large disturbance. Heavy load shedding, generator triggering

and short-circuit faults clearance are all the potential reasons. Reference (Swetapadma & Yadav, 2016) used a decision tree-based scheme for fault detection and classification during power swing within half cycle time. The decision tree algorithm is also adopted in (Jena & Samantaray, 2016) with 21 potential features extracted from phasor measurement unit (PMU) data after Kalman filter process for intelligent relaying in transmission system. A probabilistic framework is established in (Papadopoulos et al., 2018) based on the decision tree and hierarchical clustering for dynamic behavior of power systems after an occurrence of interruption. The unstable groups which may lose synchronism can be successfully detected.

Although the PMU and WAMS provide high-resolution datasets for engineers to discover patterns of normal and abnormal operation, the low probability of events that occur in power grid leads to a severe class imbalance problem. The conventional data analytics are difficult to extract the features of rare instability from massive synchrophasor measurements. Reference (Zhu et al., 2017) develops a systematic imbalance learning machine for online short-term voltage assessment. A forecasting-based nonlinear synthetic minority oversampling technique is adopted in the cost-sensitive learning algorithm to deal with the class skewness. To take full advantage of massive power grid data, the random matrix theory is introduced in (Wei et al., 2016; XUXinyi, 2016) with a high-order data-driven model to present the power system parameters and external data like meteorological information. The eigenvalue-based analysis method is proven to deal with online transient state analysis. An online monitor of instantaneous electromechanical dynamics in transmission system is presented in (Zhang et al., 2016a) based on the parallel computing and k-nearest neighbors learning algorithms. The information that indicating time-varying correlations of power generation and consumption is extracted with the proposed framework. An active learning solution is proposed in (Malbasa et al., 2017) to solve the problems for online data-driven model updating and offline training, which provide an efficient way for data sets preparation. A novel PMU-based robust state estimation method is proposed in (Zhao et al., 2016) for online state estimation of a power system under different operation conditions with the help of an adaptive weight assignment function to dynamically adjust the measurement weight according to the large disturbance revealed from PMU data. A similar framework is proposed in (Shah et al., 2017) to enable the utility company for real-time data processing. The core vector machine (CVM) is used for a two-class classification in (Wang et al., 2016b) to process the huge amount of PMU data from power grid. The CVM model is trained offline with 24 features extracted from the raw data for an online assessment evaluation for the TSA problem. The transient stability boundary of large-scale power systems is analyzed in (Lv et al., 2017b) by a statistical nonparametric regression methodology based on the critical clearing time to determine whether a steady-state condition can recover after a given fault.

### Electric device state estimation/health monitoring

As a vital component for electrical energy conversion, a failure in power transformers may cause catastrophic blackouts in power system (Reinhardt & Reinhardt, 2016). Therefore, the

life-cycle management of power transformers based on an accurate estimation attracts a lot of researches for a more stable and reliable power grid. The existing diagnosis methods for power transformers mainly focus on limited state parameters with the threshold-based diagnosis. To take information of system operation and meteorological conditions into state estimation analysis, three classical algorithms for association rule mining are discussed in (Sheng et al., 2018), namely, Apriori, AprioriTid and AprioriHybrid. The rule mining methods are combined with probabilistic graphical model for potential failure prediction.

In most commercial buildings, the building automation system (BAS) are designed and adopted to control the heating, ventilating and airing conditioning (HVAC) system to maintain proper temperature and humidity for the occupants. If the indoor smart grids can be monitored on a continuous or regular basis, a proper operation strategy may be proposed for the improvement of energy efficiency, fault diagnosis and system reliability. In (Allen et al., 2016) a novel health monitoring system is proposed by the fuzzy logic for abnormal operating condition detection. The fault signatures for various fault types are generated by the ANN classification technique.

As the rising number of aging assets in power system is becoming a potential threat to the safety operation, a lot of failure models are proposed focusing on variables of aging time or conditions. Reference (Murthy et al., 2004) proposed a failure rate model for general electric power equipment with the lifecycle data of service age, maintainer, health index taken into consideration. In order to make the best use of these data, the stratified proportional hazards model (PHM) is developed as a nonparametric regression method to process and classify the lifecycle data into multi-type recurrent events quantitatively (Qiu et al., 2016). The potential risk problem and health condition can be predicted with the help of this PHM method (Colombo et al., 1985).

### Power quality monitoring

As a worldwide issue, Electric power quality (PQ) refers to the magnitude, frequency and waveform of voltage and current in power system and highly related to the safe operation of power grid as well as the satisfaction of consumers. With the increasing application of nonlinear and power electronics based loads and generators, the harmonic distortions and instable situations frequently appears in power grid. Deep learning is successfully employed for the classification of PQ events of the electricity networks in (Balouji & Salor, 2017). Instead of sampling the voltage data of the PQ event data like the existing analysis methods, the image files of the three-phase PQ events are processed for classification by deep learning techniques. Due to the high cost for installation of advance metering devices, the conventional electromechanical analog meters still work in some residential areas and the data analytics-based PQ analysis cannot be properly utilized. Reference (Tang et al., 2015) presents a framework that collecting electricity information of from analog meters via image processing techniques. The power consumption information can then be collected to a cloud server through online data exchange. Under the consideration of balance between computation capability and the satisfactory performance of the algorithm, a compact method is presented in (Borges et al., 2016) for feature extraction from the raw data in smart grid to get information that is highly related to the field of power quality. A robust and fast processing pattern recognition algorithm is proposed in power quality events (PQE) classification is illustrated in (Ferhat et al., 2016). The features

highly correlated to the PQE are extracted with the discrete wavelet transform-entropy and basic statistical criteria for the establishment of ELM classifier.

### Topology identification

Taking the advantage of information layers in smart grid is an effective means to approach the challenges from the renewable energy sources (RES) in distribution network. The measurement, monitoring, communication and control of smart grids by advanced sensors and devices are making the complex network sensible and perceptible. The randomness of RES and uncertainty of the load are increasing the urgency and necessity for a comprehensive decision based on huge volume of data collecting and processing. The SCADA and WAMS provide voltage and power data of smart grid in near real-time sampling rate (Gungor et al., 2011) (Ghosh et al., 2013). Since the network-constrained economic dispatch problems are supposed to be solved by the real-time electricity process in a contemporary whole-sale electricity market, the potential of recovering the topology of a grid is explored with market data in (Kekatos et al., 2014). Another dynamic solution for online SG topology identification (TI) is proposed in (Babakmehr et al., 2016) which is reformulated as a sparse-recovery problem. Grapy theory and probabilistic DC optimal power flow are adopted for building the network model.

With the purpose for a greener society, the low carbon technologies (LCTs) are driven by the government by application of heat pumps, photovoltaic, electric vehicles and other smart appliances in low voltage (LV) distribution networks. Therefore, the visualization of LV networks with limited metering and data acquisition equipment attracts increasing research interests. The network load profiling based on the identification of representative load profiles of LV systems is an economical alternative method. A novel three-stage network load profiling method proposed in (Li et al., 2015a; Li et al., 2015b) aims to evaluate the capabilities of the current LV networks to accommodate the LCTs by clustering, classification and scaling. The first two stages are used to identify the load conditions of unmonitored LV systems with similar fixed data to those monitored LV substations. The contribution factor for each LV template is then determined by the cluster-wise weighted constrained regression algorithm.

### Renewable energy forecasting

The abundant and environmental friendly RES such as wind and photovoltaic energies are supposed to be the dominant energy source for the next generation of power grid. However, the randomness and intermittent characteristics are always obstacles for a large-scale utilization of RES in a stable way. To deal with such enormous challenges and get an improved dispatch planning, maintenance scheduling as well as regulation, an accurate and reliable RES forecasting approach has become the hot spot around the world (Ak et al., 2016). A data mining based method consisting of k-means and neural networks is proposed in (Wenbin & Peng, 2017). The meteorological information in historical records are used for clustering approach to classify the days into different categories. Then the bagging algorithm based neural network is trained to get the forecasting results of wind energy. Instead of using the neural network, (Ye et al., 2016) utilizes the support vector regression method to predict the wind speed with the time series historical wind speed processed by empirical mode decomposition into several intrinsic mode functions and

residue. In (Yang et al., 2015) a short-term probabilistic wind generation forecast method is presented based on the sparse Bayesian classification and Dempster-Shafer theory as a nonparametric approach. Reference (Khodayar et al., 2017) studies the ultra-short-term wind forecasting with the deep learning method through unsupervised feature learning from the unlabeled historical wind speed data. The forecasting approach of distributed solar energy systems from macro- and micro-aspects is discussed in a general way in (ZHAO et al., 2017) with clustering of capacity and location of PV system. The data-driven forecasting approach of PV diffusion is proposed based on cellular automation in microscopic analysis. By decomposing the time-series data with discrete wavelet transform, the proposed recurrent neural network (RNN) model in (Nazaripouya et al., 2016) is developed for ultra-short-term solar power prediction.

### Load forecasting

Like the RES prediction, an accurate short-term load forecasting is the essential basis for energy management, system operation and market analysis. As is mentioned in (Bunn & Farmer, 1985; Ni et al., 2016), an increase of forecasting accuracy may bring a lot of benefits and save the investments. With the emerging active role of customers in smart grid, the high efficient dynamic electricity market is also based on a good performance of electricity consumption prediction. Since electricity consumption is affected by the weather conditions to some extent, reference (Liu et al., 2018) proposed a Map/Reduce programming framework for distributed load forecasting by partitioning the geographical area according to local weather information. An extreme learning machine ensembled with a novel wavelet transform is used for electricity consumption in (Li et al., 2016a) after a conditional mutual information based feature selection, which is also used in (Ahmad et al., 2017). To overcome the volatility and uncertainty of load profiles, the recurrent neural network is adopted with a novel pooling layer to avoid overfitting problems in (Shi et al., 2017). Rather than the aggregated load forecasting, the energy consumption in a single house is usually volatile and difficult to be predicted. Driven by the recent success of deep learning, a long short-term memory recurrent neural network based framework in (Kong et al., 2017) is applied to the residential load forecasting as the latest deep learning techniques. A hidden mode Markov decision process model is developed in (Li & Jayaweera, 2015) to the forecast the customers' real-time behavior. Reference (Moreno-Munoz et al., 2016) analysis the emerging trends and challenges in the new era of using social media through mobile apps to improve their customer engagement and load forecast. Reference (Cai et al., 2017) considers the impact of social activities on the prosumers' arrangements for their generation and consumption patterns and further discuss the overall impact on the final load and the network usage.

### Load profiling

Load profiling is a way to describe the typical behavior of electric consumption, which is usually represented in time domain for load forecasting, demand-side management and capital planning (Wenhao et al., 2016; Claessens et al., 2016; Granell et al., 2015; Singh & Yassine, 2017; IMRAN et al., 2016). As an effective method for energy management, the tariff structure designed before is usually based on the type of activity, which is not able to indicate the electrical behavior in a comprehensive way (Ahmed et

al., 2017). Reference (Bo et al., 2016) utilized a two-stage clustering algorithm to classify customers according to their load curves. In the first-stage, the load patterns are clustered into different categories according to the evaluation index, and then the customers are classified according to the comprehensive load shape factors defined in the first-stage with SVM algorithm. In contrary to the time domain analysis (Al-Otaibi et al., 2016), the DFT method is adopted in (Zhong & Tam, 2012) to discover the information of customers' behavior, which can be accurately reconstructed using limited frequency components and still satisfy the strict requirements. The residential electricity consumption usually can be divided into three parts: fixed, regulable and deferrable loads, which is the theoretical basis for the optimal energy management of the demand response (DR) mechanism. DR is used to initiate a change in the customers' consumption or feed-in pattern with an incentive from costs or ecological information. Reference (Li et al., 2017) utilizes the spectral domain analysis methods DWT and DFT to decompose smart metering data with the extracted coefficients. Results show that DWT performs better than DFT in individual level while DFT is more suitable to be used in the analysis at a highly aggregated level. A learning based DR strategy combining data analytics and optimization is developed for regulatable loads focusing on the residential HVAC (Zhang et al., 2016b). Because when the customers' behavior is obtained, an optimal DR technique for household HVAC unit can be designed based on weather prediction, day-ahead electricity price. Reference (Jindal et al., 2016) takes the advantage of the social networking to minimize the peak power consumption of the electrical appliances by proposing a "family plan" approach which leverages the social network topology and statistical energy usage patterns of the users.

To better understand the information behind the stochasticity and irregularity of residential energy consumption, an in-depth analysis is presented in (Grindrod, 2016) with a finite mixture model-based clustering technique. The self-organizing maps (SOM) as a type of ANN is used in (Verdú et al., 2006) to reduce the dimension of collected raw data for load pattern extraction. The frequency-domain data analytics in the SOM shows a superiority over the time-domain data with a higher accuracy in new customer classification. As one of the main tasks of load profiling, a better understanding of the flexibility of customers' electricity consumption is the basis for DR, which can be used to release the pressure of distribution system in terms of thermal and voltage constrains. A multi-resolution analysis method based on wavelet analysis is proposed in (Li et al., 2016b) to extract spectral and time-domain features of load data. Different permutations of typical load profiles provide a more flexible load profiling with a reduction of computation. With the popularization of electric vehicles (EVs), learning the charging load patterns of them is becoming a key step for the stability of power grids. An unsupervised clustering algorithm is used in (Munshi & Mohamed, 2018) to extract the pattern of EV charging loads with only the real power measurements. Furthermore, the flexibility of the collective EV charging demand is analyzed with Bayesian maximum likelihood. References (Tong et al., 2016; Wang et al., 2017b) focus on the problem brought by the huge load profile data with the popularity of smart meters installed at the household level, which poses challenges to the communication and storage of measurement data as well as the vital information extraction from massive records. K-SVD sparse representation technique is used to decompose the load profiles into several partial usage patterns for a linear SVM based method to recognize the type of customers.

### Load disaggregation

Load disaggregation is also called non-intrusive load monitoring (NILM), aiming to segregate the overall load profiles at household level into the energy consumption of individual appliances. Unlike direct appliance monitoring framework, the NILM from only one smart meter installed in the house is easier to be accepted by the customers (Liang et al., 2010a; Liang et al., 2010b; Gillis et al., 2016). Since different types of the household electric appliances have different potential to be involved in the DR program, the appliance-level load profiles allow the utilities to understand the customers' behavior better and helps to develop a more energy efficient strategy. The early techniques for NILM are mainly based on the detection of "edge" in power signal to indicate the state "on" or "off" of a known device (Sultanem, 1991). The more effective and complex appliance signatures are then proposed with the harmonics computation of steady-state power or current (Berges et al., 2009; Lee et al., 2004). The hidden Markov models (HMMs) are adopted in (Kong et al., 2018) with the segmented integer quadratic constraint programming to disaggregate the household power profile at an average frequency of 0.3 Hz into the appliance level. In (Henao et al., 2017) a NILM approach based on the subtractive clustering the maximum likelihood classifier is proposed for a low-sampling-rate date set of 1 Hz sampling rate. The appliances are modeled as ON/OFF states in this event-based load disaggregation algorithm. As a single channel blind source separation problem, the dictionary learning based approaches can be used in NILM. A deep learning approach with multiple layers of dictionaries trained for each device as "deep sparse coding" is utilized in (Singh & Majumdar, 2017; Zico Kolter et al., 2010). Compared with HMM, the latter method is not suitable for real-time application. By combining the decision tree and nearest-neighbor algorithms, the semi-supervised machine learning is applied to the NILM problem in (Gillis & Morsi, 2017) with the signal features extracted by matching a set of net wavelets to the load classes.

### Non-technical loss detection

The nontechnical loss (NTL), which is probably caused by the electrical theft or errors in accounting, is one of the prominent concerns that have plagued the power system utilities for a long time (Leung, 2016; Zhan et al., 2016; Non-Cooperative Game Model Applied to an Advanced Metering Infrastructure for Non-Technical Loss Screening in Micro-Distribution Systems, 2014; Guerrero et al., 2018). According to the survey published by Northeast Group, LLC, the loss caused by electricity theft reached more than $89.3 billion in the world every year (PR Newswire, 2014). Furthermore, large scale electricity fraudulent behavior may cause severe imbalance problems in power system. Therefore, the effective framework to detect the NTL in the complex power grid has appealed many research interests. A comprehensive top-down scheme based on DT and SVM is proposed in (Jindal et al., 2016). DT is trained with various features including heavy appliances, number of persons, weather conditions to get the expected value of electricity consumption for the customer during a particular time. Then the calculated consumption along with other features are fed to the SVM classifier which is already trained based on the collected dataset to determine whether the customer's behavior is normal or fraud. In (Zanetti et al., 2017) the fraud detection is triggered when a discrepancy is detected between energy supplied from the power system and

collected information from smart meters. The anomalies in consumption patterns are discovered with the fuzzy clustering algorithm.

### Open issues for the application of big data analytics in smart grids

Even though there are increasing researches on the big data analytics in smart grids, the deployed applications are few. There are still many open issues needed to be addressed before the techniques can create implications in reality.

With the fast deployment of smart meters and advanced sensors, huge amount of data with multiple types and structures from deference sources with a variety of protocols are generated every second. However, the lack of standard data format for the information software and database structures, as well as the issue of interoperability of different information and communication systems deployed in the smart grids, make it complicated and difficult to obtain data for real application. The traditional way of isolated storage of the data in various systems also increases the barrier for data sharing among applications.

As a conventionally sensitive industry, most of the data generated in the smart grid are considered as confidential or related with privacy issues; therefore, it is impractical for researchers to conduct highly relevant studies which can be smoothly transferred later on into deployment. Thus, most of the researches are still about the algorithms which are tested with ideal data, and hence stay in the Ivory tower.

In addition, due to the lack of strategic vision, top design of application, large investment in reality, combined with the short-sighted recognition of the value of the data, the applications of big data in real systems are growing very slow. Even though, the majority of utility companies showed great interests in the big data analytics and their application in their business, they are still waiting to see convincing results before they are willing to put more efforts and investment.

Last but not least, the big data analytics in smart grids is a comprehensive and complicated field, which does not only depend on the mathematic algorithms or techniques, it also depends on the operation of the systems, the behaviors of vast number of autonomous users, the ICT technologies, the expertise of the field, etc. Therefore, it needs the synergy among experts from different fields if we would like to see the benefits of it in the smart grids.

## Conclusion

In this article, the big data in smart grid and the corresponding state-of-the-art analysis methods have been reviewed and discussed. The data which may contain valuable information are collected from smart meters installed in the power system, electricity market, GIS, meteorological information system, social media, and so on. The purpose of advanced ICT technology in power system is to associate the traditional physical parameters in power system to the external variables to discover potential regulations and scientific problems. Eleven applications of data analytics mentioned in the paper are nearly involved in every aspect of smart grids, including the operation, maintenance, load/output forecasting, protection as well as fault detection and location. After extracting the useful features from raw information with the background knowledge of electrical engineering, typical data analytics methods, such as neural network, k-means, and support vector machine, could be widely applied. Secure and efficient operation

strategies as well as optimal business decisions are supposed to be made with the data analytics from a more unified view.

With more advanced ICT technologies applied in power system, the fast and efficient data analytics framework for huge volume of data would become a challenging requirement. Moreover, the cyber security and privacy protection could become as important as a relay protection in power system. Even though the interactive communication with customers provides a potential solution for more accurate demand response, it also increases the difficulties in consumption behavior analysis at the same time. A secure and high-performance data analytics platform would be crucial for the social welfare and power companies' interests in the future. As the application of data analytics in smart grids is a comprehensive and complicated field, involving mathematics, ICT technologies, computer science, electrical engineering, etc., thus, it needs the synergy among experts from different fields as well as the strategic visions for the top designs.

**Funding**
This paper is completed under no funding.

**Availability of data and materials**
Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

**Authors' contributions**
EB carried out the smart grid studies, participated in the concepts of big data in power system and the structure of the paper. TH participated in the frontier technologies in smart grid and drafted part of the manuscript. YZ participated in the data analysis application survey in smart grid and drafted the manuscript. All authors read and approved the final manuscript.

**Competing interests**
The authors declare that they have no competing interests.

## Publisher's Note



Received: 27 January 2018 Accepted: 11 June 2018
Published online: 13 August 2018

**References**

Ahmad A, Javaid N, Guizani M, Alrajeh N, Khan ZA (Oct. 2017) An accurate and fast converging short-term load forecasting model for industrial applications in a smart grid. IEEE Transactions on Industrial Informatics 13(5):2587–2596

Ahmed N, Levorato M, Li GP (2017) Residential consumer-centric demand side management. IEEE Transactions on Smart Grid. https://doi.org/10.1109/TSG.2017.2661991

Ak R, Fink O, Zio E (2016) Two machine learning approaches for short-term wind speed time-series prediction. IEEE Transactions on Neural Networks and Learning Systems 27(8):1734–1747

Alam Mollah Rezaul, Kashem M. Muttaqi, Abdesselam Bouzerdoum. Evaluating the effectiveness of a machine learning approach based on response time and reliability for islanding detection of distributed generation. IET renewable power generation (volume: 11, Issue: 11, 2017)

Allen WH, Rubaai A, Chawla R (May-June 2016) Fuzzy neural network-based health monitoring for HVAC system variable-air-volume unit. IEEE Trans Ind Appl 52(3):2513–2524

Al-Otaibi R, Jin N, Member IEEE, Wilcox T, Flach P (April 2016) Feature construction and calibration for clustering daily load curves from smart-meter data. IEEE Transactions on Industrial Informatics 12(2):645–654

Andalib-Bin-Karim C, Liang X, Khan N, Zhang H (2017) Determine Q-V characteristics of grid-connected wind farms for voltage control using a data-driven analytics approach. IEEE Trans Ind Appl 53(5)

Babakmehr M, Simões MG, Wakin MB, Harirchi F (2016) Compressive sensing-based topology identification for smart grids. IEEE Transactions on Industrial Informatics 12(2):532–543

Bahmanyar A, Jamali S, Estebsari A, Pons E, Bompard E, Patti E, Acquaviva A (2016) Emerging smart meters in electrical distribution systems: opportunities and challenges. In: 24th Iranian Conference on Electrical Engineering (ICEE). Shiraz, Iran

Baimel D, Tapuchi S, Baimel N (2016) Smart grid communication technologies- overview, research challenges and opportunities. International Symposium on Power Electronics, Electrical Drives, Automation and Motion (SPEEDAM), 22-24 June 2016, Anacapri, Italy

Balouji E, Salor O (2017) Classification of power quality events using deep learning on events images. 3rd International Conference on Pattern Recognition and Image Analysis (IPRIA), 19-20 April 2017, Shahrekord, Iran

Bauman K, Tuzhilin A, Zaczynski R (2017) Using social sensors for detecting emergency events: a case of power outages in the electrical utility industry. ACM Transactions on Management Information Systems 8(2–3)

Berges M, Goldman E, Matthews HS, Soibelman L (2009) Learning systems for electric comsumption of buildings. In: In ASCI International Workshop on Computing in Civil Engineering
Big Data analytics and energy consumption. (2016) Available: http://www.lavastorm.com/blog/2012/04/09/big-data-analytics-and-energy-consumption/
Bo P, Wan C, Dong S, Lin J, Song Y, Yi Z, Xiong J (2016) A Two-stage Pattern Recognition Method for Electric Customer Classification in Smart Grid. In: 2016 IEEE International Conference on Smart Grid Communications (SmartGridComm), Sydney, vol. 12
Borges FAS, Fernandes RAS, Silva IN, Silva CBS (April 2016) Feature extraction and power quality disturbances classification using smart meters signals. IEEE Transactions on Industrial Informatics 12(2):824–833
Bunn DW, Farmer ED (1985) ComparativeModels for electrical load forecasting. Wiley, New York
Cai Y, Huang T, Bompard E, Cao Y, Li Y (2017) Self-Sustainable Community of Electricity Prosumers in the Emerging Distribution System. IEEE Transactions on Smart Grid, vol 8, no. 5, pp. 2207–2216
Cai Y, Chow M-Y (2009) Exploratory analysis of massive data for distribution fault diagnosis in smart grids. IEEE Conference on Power & Energy Society General Meeting, July
CEN-CENELEC-ETSI Smart Grid Working Group Reference Architecture (2012) Reference architecture for the smart grid. Tech Rep
Cheng Y, Chen K, Sun H, Zhang Y, Tao F (2018) Data and knowledge mining with big data towards smart production. Journal of Industrial Information Integration 9:1–13
Chunming T, Xi H, Shuai Z, Jiang F (2017) Big data issues in smart grid – a review. Renew Sust Energ Rev 79:1099–1107
Claessens BJ, Vrancx P, Ruelens F (2016) Convolutional neural networks for automatic state-time feature extraction in reinforcement learning applied to residential load control. IEEE Transactions on Smart Grid
Colombo AG, Costantini D, Jaarsma RJ (1985) Bayes nonparametric estimation of time-dependent failure rate. IEEE Trans Rel 34(2):109–112
Cui Q, El-Arroudi K, Jo'os G'e (2017) An Effective Feature Extraction Method in Pattern Recognition Based High Impedance Fault Detection. In: 2017 19th International Conference on Intelligent System Application to Power Systems (ISAP), San Antonio, 17-20 Sept. 2017
De Santis E, Livi L, Sadeghian A, Rizzi A (2015) Modeling and recognition of smart grid faults by a combined approach of dissimilarity learning and one-class classification. In: In Neurocomputing, vol 170, pp 368–383 ISSN 0925-2312
De Santis E, Rizzi A, Sadeghian A (2017) A Learning Intelligent System for Classification and Characterization of Localized Faults in Smart Grids. 2017 IEEE Congress on Evolutionary Computation (CEC), San Sebastian, 5-8 June 2017
Di Zhua TL, Zhang J (2018) Unsupervised tip-mining from customer reviews. Decis Support Syst 107:116–124
Dimitrovska T, Rudež U, Mihalič R (2017) Fast contingency screening based on data mining. In: IEEE EUROCON International Conference on Smart Technologies, Ohrid, 6-8 July 2017
Emani CK, Cullot N, Nicolle C (2015) Understandable Big Data: A survey. Computer Science Review 17:70–81
Executive Office of the President (2013) Economic benefits of increasing electric grid resilience to weather outages. In: USA
Fan C, Xiao F, Li Z, Wang J (2018) Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: a review. In: Energy and Buildings, vol 159, pp 296–308
Ferhat UÇAR, Ömer Faruk ALÇİN, Beşir DANDIL, Fikret ATA (2016) Machine Learning based Power Quality Event Classification using Wavelet_Entropy and Basic Statisticsal Features. In: 2016 21st International Conference on Methods and Models in Automation and Robotics (MMAR), Miedzyzdroje, 29 Aug.-1 Sept.2016
Ghosh D, Ghose T, Mohanta DK (Aug. 2013) Communication feasibility analysis for smart grid with phasor measurement units. IEEE Trans. Ind. Informat. 9(3):1486–1496
Gillis JM, Alshareef SM, Morsi WG (2016) Nonintrusive load monitoring using wavelet design and machine learning. IEEE Transactions on Smart Grid 7(1):320–328
Gillis JM, Morsi WG (Nov. 2017) Non-intrusive load monitoring using semi-supervised machine learning and wavelet design. IEEE Transactions on Smart Grid 8(6):2648–2655
Granell R, Axon CJ, Wallom DCH (Nov. 2015) Impact of raw data temporal resolution using selected clustering methods on residential electricity load profiles. IEEE Trans Power Syst 30(6):3217–3224
Guerrero JI, Monedero I, Biscarri F, Biscarri J, Millán R, León C (2018) Non-technical losses reduction by improving the inspections accuracy in a power utility. IEEE Trans Power Syst, vol. 33, pp. 1209-1218
Gungor V et al (Sep. 2011) Smart grid technologies communications technologies and standards. IEEE Trans Ind Informat 7(4):529–539
Günther WA, Rezazade Mehrizi MH, Huysman M, Feldberg F (2017) Debating big data: a literature review on realizing value from big data. J Strateg Inf Syst 26:191–209
Haben S, Singleton C, Grindrod P (Jan. 2016) Analysis and clustering of residential customers energy behavioral demand using smart meter data. IEEE Transactions on Smart Grid 7(1):136–144
Hashemi F, Mohammadi M, Kargarian A (2017) Islanding detection method for microgrid based on extracted features from differential transient rate of change of frequency. IET Generation, Transmission & Distribution 11(4):891–904
He C, Lin G, Mo W (2016) A method for transient stability assessment based on pattern recognition. International Conference on Smart Grid and Clean Energy Technologies (ICSGCE), 19-22 Oct. 2016
He M, Zhang J, Vittal V (Nov. 2013) Robust online dynamic security Assesment using adaptive ensemble decision-tree learning. IEEE Trans Power Syst 28(4):4089–4098
Henao N, Agbossou K, Kelouwani S, Dubé Y, Fournier M (2017) Approach in nonintrusive type I load monitoring using subtractive clustering. IEEE Transactions on Smart Grid 8(2):812–821
Imran K, Joshua Zhexue H, Md Abdul Masud, Qingshan J (2016) Segmentation of factories on electricity consumption behaviors using load profile data. IEEE Access 4:8394–8406
Jena MK, Samantaray SR (2016) Data-mining-based intelligent differential relaying for transmission lines including UPFC and wind farms. IEEE Transactions on Neural Networks and Learning Systems 27(1):8–17
Jiang H, Dai X, Gao DW, Zhang JJ, Zhang Y, Muljadi E (Sept. 2016) Spatial-temporal Synchrophasor data characterization and analytics in smart grid fault detection identification and impact casual analysis. IEEE Transactions on Smart Grid 7(5):2525–2536

Jindal A, Dua A, Kaur K, Singh M, Kumar N, Mishra S (2016) Decision tree and SVM-based data analytics for theft detection in smart grid. IEEE Transactions on Industrial Informatics 12(3):1005–1016
Jiye Q, Zhixiang J, Mengjie S et al (2015) Scenario analysis and application research on big data in smart power distribution and consumption systems. Proceedings of the CSEE 35(8):1829–1836
Jokar P, Arianpoo N, Leung VCM (2016) Electricity theft detection in AMI using customers' consumption patterns. IEEE Transactions on Smart Grid 7(1):216–226
Kaisler S, Amnour F, Alberto J (2012) "Big data: issues and challenges moving forward", 46th IEEE international conference on system science, Wailea, Maui, HI, USA, 7-10 Jan. 2013
Kar S, Samantaray SR, Dadash Zadeh M (2017) Data-mining model based intelligent differential microgrid protection scheme. IEEE Syst J 11(2):1161–1169
Kekatos V, Giannakis GB, Baldick R (2014) Grid topology identification using electricity prices. In: Proc. IEEE Power Energy Soc. Gen. Meeting. National Harbor, MD, USA, pp 1–5
Keyan L, Wanxin S, Dongxia Z et al (2015) Big data application requirements and scenario analysis in smart distribution network. Proceedings of the CSEE 35(2):287–293
Khodayar M, Kaynak O, Khodayar ME (Dec. 2017) Rough deep neural architecture for short-term wind speed forecasting. IEEE Transactions on Industrial Informatics 13(6):2770–2779
Kong W, Dong ZY, Jia Y, Hill DJ, Xu Y, Zhang Y (2017) Short-term residential load forecasting based on LSTM recurrent neural network. IEEE Transactions on Smart Grid (Early Access) https://doi.org/10.1109/TSG.2017.2753802
Kong W, Dong ZY, Ma J, Hill DJ, Zhao J, Luo F (2018) An extensible approach for non-intrusive load disaggregation with smart meter data. IEEE Transactions on Smart Grid 9(4):3362–3372
Lee W, Fung G, Lam H, Chan F, Lucente M (2004) Exploration on load signatures. International Conference on Electrical Engineering (ICEE)
Li D, Jayaweera SK (Dec. 2015) Machine-learning aided optimal customer decision for an interactive smart grid. IEEE Syst J 9(4):1529–1540
Li R, Gu C, Li F, Shaddick G, Dale M (2015a) Development of low voltage network Templates_Part I_Substation clustering and classification. IEEE Trans Power Syst 30(6)
Li R, Gu C, Li F, Shaddick G, Dale M (2015b) Development of low voltage network templates—part II_ peak load estimation by Clusterwise regression. IEEE Trans Power Syst 30(6)
Li R, Li F, Smith ND (Nov. 2016b) Multi-resolution load profile clustering for smart metering data. IEEE Trans Power Syst 31(6):4473–4482
Li R, Li F, Smith ND (June 2017) Load characterization and low-order approximation for smart metering data in the spectral domain. IEEE Transactions on Industrial Informatics 13(3):976–984
Li S, Wang P, Goel L (May 2016a) A novel wavelet-based ensemble method for short-term load forecasting with hybrid neural networks and feature selection. IEEE Trans Power Syst 31(3):1788–1798
Liang Jian, Simon K. K. Ng, Gail Kendall, and John W. M. Cheng. Load Signature Study-Part I: Basic Concept Structure and Methodology. IEEE Transactions on Power Delivery ( 25: 2, 2010): 551–560
Liang J, Ng SKK, Kendall G, Cheng JWM (2010b) Load signature study-part II: disaggregation framework simulation and applications. IEEE Transactions on Power Delivery 25(2):561–569
Lim Ee-Peng, Jaideep Srivastava, Satya Prabhakar, James Richardson, Entity identification in database integration, in information sciences, 89:1–2, 1996;1–38, ISSN 0020-0255
Liu C, Sun K, Rather ZH, Chen Z, Bak CL, Thøgersen P, Lund P (2014) A Systematic Approach for Dynamic Security Assessment. IEEE Transactions on Power Systems 29(2):717–730
Liu D, Zeng L, Li C, Ma K, Chen Y, Cao Y (2018) A distributed short- term load forecasting method based on local weather information. IEEE Syst J vol. 12, pp. 208-215
Lv J, Pawlak M, Annakkage UD (2017a) Prediction of the transient stability boundary based on Nonparameteric additive modeling. IEEE Trans Power Syst 32(6):4362–4369
Lv Z, Song H, Basanta-Val P, Steed A (2017b) Analytics MJN-GBD State of the art, challenges and future research topics. IEEE Transactions on Industrial Informatics 13(4):1891–1899
Madhumita P, Tajane S, Indi B (2016) Assessment of system vulnerability for smart grid applications. IEEE International Conference on Engineering and Technology (ICETECH)
Malbasa V, Zheng C, Chen P-C, Popovic T, Kezunovic M (Nov. 2017) Voltage stability prediction using active machine learning. IEEE Transactions on Smart Grid 8(6):3117–3124
Mishra DP, Samantaray SR, Joos G (2016) A combined wavelet and data-mining based intelligent protection scheme for microgrid. IEEE Transactions on Smart Grid 7(5):2295–2304
Moreno-Munoz A, Bellido-Outeirino FJ, Siano P, Gomez-Nieto MA (2016) Mobile social media for smart grids customer engagement: emerging trends and challenges. Renew Sust Energ Rev 53:1611–1616
Munshi AA, Mohamed YA-RI (2018) Extracting and defining flexibility of residential electrical vehicle charging loads. IEEE Transactions on Industrial Informatics vol 14, pp. 448-461
Murthy DNP, Bulmer M, Eccleston JA (Dec. 2004) Weibull model selection for reliability modeling. Rel Eng Syst Safety 86(3):257–267
Nazaripouya H, Wang B, Wang Y, Chu P, Pota HR, Gadh R (2016) Univariate Time Series Prediction of Solar Power Using a Hybrid Wavelet-ARMA-NARX Prediction Method. In: 2016 IEEE/PES Transmission and Distribution Conference and Exposition (T&D), Dallas
Ni D, Benoit C, Foggia G, Bésanger Y, Senior Member IEEE, Wurt F (Jan. 2016) Neural network-based model Design for Short-Term Load Forecast in distribution systems. IEEE Trans Power Syst 31(1):72–81
Non-Cooperative Game Model Applied to an Advanced Metering Infrastructure for Non-Technical Loss Screening in Micro-Distribution Systems. IEEE Transactions on Smart Grid, vol. 5, no. 5, 2014: 2468–2469
Papadopoulos PN, Guo T, Milanović JV (2018) Probabilistic framework for online identification of dynamic behavior of power systems with renewable generation. IEEE Trans Power Syst , vol. 33, pp. 45-54
PR Newswire. (2014). "World Loses $89.3 Billion to Electricity Theft Annually, $58.7 Billion in Emerging Markets." [Online]. Available: http://www.prnewswire.com/news-releases/world-loses-893-billion-to-electricity-theft-annually-587-billion-in-emerging-markets-300006515.html, Accessed on: Jul. 2015

Qiu J, Wang H, Lin D, He B, Zhao W, Wei X (2016) Nonparameteric Regression-based Failure Rate Model for Electric Power Equipment Using Lifecycle Data. In: 2016 IEEE/PES Transmission and Distribution Conference and Exposition (T&D), Dallas
Reinhardt A, Reinhardt D (2016) Detecting Anomalous Electrical Appliance Behavior based on Motif Transition Likelihood Matrices. In: 2016 IEEE International Conference on Smart Grid Communications (SmartGridComm), Sydney, NSW, Australia, Sydney, NSW, Australia
Roya A, Cruz a RMO, Sabourina R, Cavalcanti GDC (2018) A study on combining dynamic selection and data preprocessing for imbalance learning. Neurocomputing 286:179–192
Sagiroglu S, Terzi R, Canbay Y, Colak I (2016) Big Data Issues in Smart Grid Systems. In: 2016 IEEE International Conference on Renewable Energy Research and Applications (ICRERA), Birmingham, pp 20–23
Shah Z, Anwar A, Mahmood AN, Tari Z, Zomaya AY (2017) A Spatio-temporal Data Summarization Paradigm for Real-time Operation of Smart Grid. IEEE Transactions on Big Data PP(99)
Sheng G, Hou H, Jiang X, Chen Y (2018) A novel association rule mining method of big data for power transformer state parameters based on probabilistic graph model. IEEE Transactions on Smart Grid, vol. 9, pp. 695-702
Shi H, Xu M, Li R (2017) Deep learning for household load forecasting – a novel pooling deep RNN. IEEE Transactions on Smart Grid, https://doi.org/10.1109/TSG.2017.2686012
Singh S, Yassine A (2017) Mining energy consumption behavior patterns for households in smart grid. IEEE Transactions on Emerging Topics in Computing
Singh S, Majumdar A (2017) Deep Sparse Coding for Non-intrusive Load Monitoring. IEEE Transactions on Smart Grid
Siryani J, Tanju B, Eveleighi TJ (2017) A machine learning decision-support system improves the internet of things' smart meter operations. Accident Analysis and Prediction, volume 4:1056–1066
SmartGrids European Tech. Platform, Strategic Deployment Document for Europe's Electricity Networks of the Future 6 (2010) [hereinafter E.U. SmartGrids SDD]
Sultanem F (1991) Using appliance signatures for monitoring residential loads at meter panel level. IEEE Transaction on Power Delivery 6(4)
Sun H, Wang Z, Wang J, Huang Z, Carrington NL, Liao J (2016) Data-driven power outage detection by social sensors. IEEE Transactions on Smart Grid 7(5):2516–2524
Swetapadma A, Yadav A (2016) Data-mining-based fault during power swing identification in power transmission system. IET Science, Measurement & Technology 10(2):130–139
Tang Y, Ten C-W, Wang C, Parker G (2015) Extraction of energy information from analog meters using image processing. IEEE Transactions on Smart Grid 6(4)
Teng Z, Yan Z, Dongxia Z (2014) Application Technology of big Data in smart distribution grid and its Prospect analysis. Power System Technology 38(12):3305–3312
Tong X, Kang C, Xia Q (2016) Smart metering load data compression based on load feature identification. IEEE Transactions on Smart Grid 7(5):2414–2422
Verdú SV, García MO, Senabre C, Marín AG, Franco FJG (2006) Classification filtering and identification of electrical customer load patterns through the use of self-organizing maps. IEEE Trans Power Syst 21(4):1672–1682
Wang B, Member BF, Wang Y, Liu H, Liu Y (2016b) Power system transient stability assessment based on big data and the Core vector machine. IEEE Transactions on Smart Grid 7(5):2561–2570
Wang Jian, Xiaofu Xiong, Ning Zhou, Zhe Li, Wei Wang. Early warning method for transmission line galloping based on SVM and AdaBoost bi-level classifiers. IET Generation, Transmission & Distribution (Volume: 10, Issue: 14, 11 4 2016a): 3499–3507
Wang X, McArthur S, Strachan S, Kirkwood J, Paisley B (2017a) A data analytic approach fault diagnosis and prognosis for distribution automation. IEEE Transactions on Smart Grid
Wang Y, Chen Q, Kang C, Xia Q, Luo M (2017b) Sparse and redundant representation-based smart meter data compression and pattern extraction. IEEE Trans Power Syst 32(3):2142–2151
Wei L, Zhang D, Wang X, Liu D, Wu Q. Power System Transient Stability Analysis Based on Random Matrix Theory. Proceedings of the CSEE,36 ;18 pp: 4854–4863, 2016
Wenbin W, Peng M (2017) A Data Mining Approach Combining K-Means Clustering with Bagging Neural Network for Short-term Wind Power Forecasting. IEEE Internet of Things Journal 4(4):2327–4662
Wenhao P, Zhe D, Yanping Z, Jun L (2016) An analytical method for intelligent electricity use pattern with demand response. In: 2016 China International Conference on Electricity Distribution (CICED), Xi'an
Xu X, He X, Ai Q, Qiu Caiming. A Correlation Analysis Method for Operation Status of Distribution Network Based on Random Matrix Theory. Power System Technology, Vol. 40 No. 3, pp: 781–790, 2016
Yang M, Lin Y, Han X (2015) Probabilistic Wind Generation Forecast Based on Sparse Bayesian Classification and Dempster-Shafer Theory. In: 2015 IEEE Industry Applications Society Annual Meeting, Addison
Ye R, Suganthan PN, Srikanth N (2016) A novel empirical mode decomposition with support vector regression for wind speed forecasting. IEEE Transactions on Neural Networks and Learning Systems 27(8):1793–1798
Yu C (2002) Pan Heping. Business intelligence and its key technology Application Research of Computers 9:14–16
Zanetti M, Jamhour E, Pellenz M, Penna M, Zambenedetti V, Chueiri I (2017) A tunable fraud detection system for advanced metering infrastructure using short-lived patterns. IEEE Transactions on Smart Grid
Zhan T-S, Chen S-J, Kao C-C, Kuo C-L, Chen J-L, Lin C-H (2016) Non-technical loss and power blackout detection under advanced metering infrastructure using a cooperative game based inference mechanism. IET Gener Transm Distrib 10(4):873–882
Zhang D, Li S, Sun M, O'Neill Z (July 2016b) An optimal and learning-based demand response and home energy management system. IEEE Transactions on Smart Grid 7(4):1790–1801
Zhang J, Chung CY, Wang Z, Zheng X (April 2016a) Instantaneous electromechanical dynamics monitoring in smart transmission grid. IEEE Transactions on Industrial Informatics 12(2):844–852
Zhang Y, Yan X, Dong ZY, Zhao X, Wong KP (Oct. 2017) Intelligent early warning of power system dynamic insecurity Risk_Toward optimal accuracy-earliness tradeoff. IEEE Transactions on Industrial Informatics 13(5): 2544–2554
Zhang Zhen. Smart Grid in America and Europe: Similar Desires, Different Approaches. Public Utilities Fortnightly, 149, 1, 2011

Zhao J, Zhang G, Das K, Korres GN, Manousakis NM, Sinha AK, He Z (2016) Power System Real-Time Monitoring by Using PMU-based Robust State Estimation Method. IEEE Transactions on Smart Grid 7(1):300–309
Zhao T, Ziqiang Z, Yan Z, Ping L, Yingjie T. Spatio-temporal analysis and forecasting of distributed PV systems Diffusion_a Case study of shanghai using A data-driven approach. IEEE Access 5: 5135–5148, 2017
Zhong S, Tam K-S (May 2012) A frequency domain approach to characterize and Anlyze load profiles. IEEE Trans Power Syst 27(2):857–865
Zhu L, Chao L, Dong ZY, Hong C (2017) Imbalance learning machine-based power system short-term voltage stability assessment. IEEE Transactions on Industrial Informatics 13(5):2533–2543
Zhu Ting, Sheng Xiao, Qingquan Zhang, Yu Gu, Ping Yi, and Yanhua Li, "Emergent Technologies in big Data Sensing: a survey", International Journal of Distributed Sensor Networks, Volume 2015, Article ID 902982
Zico Kolter J, Batra S, Ng AY (2010) Energy Disaggregation via Discriminative Sparse Coding. In: NIPS'10 Proceedings of the 23rd International Conference on Neural Information Processing Systems, vol 1, Vancouver, pp 1153–1161
Zikopoulos P, C. Eaton, Understanding big data: analytics for Enterprise class Hadoop and streaming data, McGraw-hill Education, 2011