Journal of
Advanced Manufacturing Science and Technology
jamst2020@126.com
www.jamstjournal.com

Research and application progress of data mining technology in electric power system

Fangwei NING, Yan SHI*, Yishu CAI, Weiqing XU

School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China;

Received 25 April 2021; revised 10 May 2021; accepted 25 May 2021; Available online 22 June 2021

Abstract: With the rapid development of computer technology and the improvement of intelligent technologies in electric power engineering, the volume of data has increased exponentially. Data mining technology can be utilized to search information hidden in the huge amounts of data, and then the data can be transformed into useful knowledge to promote the development of electric power technology. In order to be acquainted with the research and application progress of data mining technology in electric power engineering, several major data mining algorithms are introduced in this paper, including ANN (Artificial Neural Network) algorithm, SVM (Support Vector Machine) algorithm, decision tree algorithm, K-means algorithm, NBC (Naive Bayesian Classification) algorithm and Apriori algorithm. And then, the methods of data mining technology in prediction, classification, clustering and association rules analysis are explained in detail in this engineering, which are combined with the electricity price prediction, power load forecasting, fault type identification, system state classification, power generation side association rules, power grid operation data association analysis. At last, this technology in electric power engineering is summarized and an expectation for the future development is provided.

Keywords: Data mining; Electric power; Data cleaning; K-means; Clustering;

1. Introduction

With the rapid development of computer technology and the improvement of the intelligent technologies, engineering measurement data, sales data, trade data, medical data, financial data and other human accumulated data have increased exponentially. "Drowning in information, but thirsty for knowledge " [1], as John Naisbitt mentioned in his book, facing the immensity of the ocean of data, we are looking for effective methods that can automatically analyze, classify and collect data, automatically discover trends and mark anomalies in data. Under such situation, data mining technology is born [2-6].

2. Concept of data mining technology

Data mining is the core step of KDD (Knowledge Discovery in Database). It is the process of revealing meaningful relationships, trends and patterns by analyzing a large amount of data. It involves statistics, machine learning, artificial intelligence, data visualization, database technology, data visualization and other disciplines[7]. Data mining is process aimed at acquiring knowledge, including data cleansing, data mining, result expression and interpretation [8]. As shown in Fig.1, data preparation mainly includes original quality inspection (data noise and loss) and correlation analysis among attributes. This process also includes 3 sub steps: data preparation, data mining and result expression and interpretation, which can distinguish the most valuable information, draw the inner link between things, provide decision support for decision-makers and even predict the future [9-11].

2. Main methods of data mining technology

2.1. Data cleaning

The quality of the data is the key factor. A large number of uncertain data need to be cleaned, including inaccuracy and uncertainty of raw data. The main purpose of data cleaning technology is to remove erroneous and inconsistent data. It first appeared in the United States, starting from the correction of the wrong social insurance numbers [12]. With the development of data mining technology, data cleaning technology has been stimulated and mainly involves data recognition, detection, elimination and integration [13].

2.1.1. Detection of abnormal data

According to different attributes, data can be divided into numeric types and character types. For the numerical type, the statistical method of calculating the mean and standard deviation of the field is used to compare the confidence interval of the set to detect and eliminate the abnormal data [13], which is the common way; for character type, the common methods are: edit distance method, incremental recognition method and LOF (Local Outlier Factor) algo-

* Corresponding author. *E-mail address:* nfangwei@163.com(Yan SHI)

Peer review under responsibility of Editorial Committee of JAMST

DOI: 10.51393/j.jamst.2021007

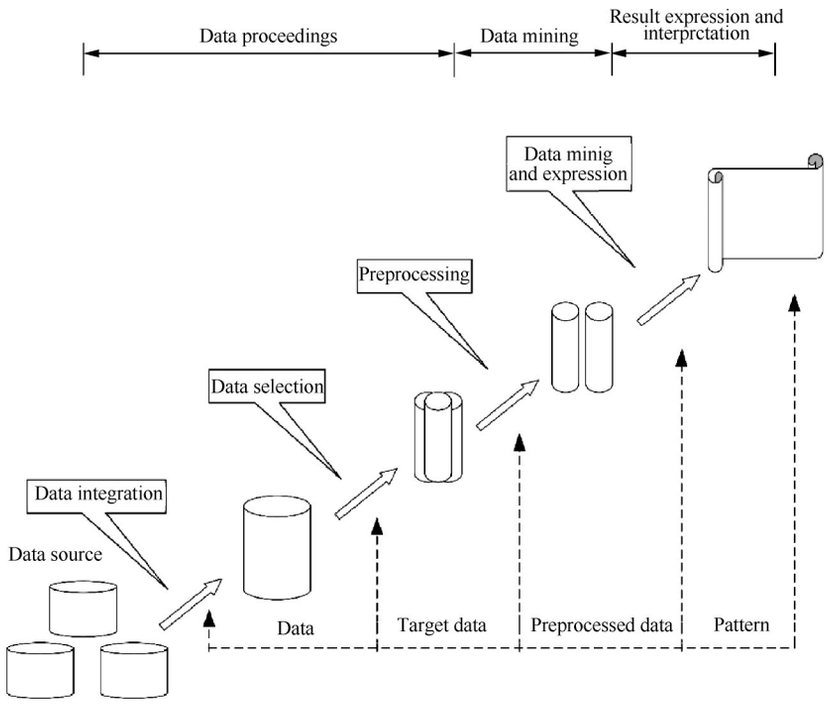

Fig. 1 Data mining process.

rithm [14, 15], which is also the most widely used method at present.

2.1.2. Elimination of similar duplicate records

Identifying and eliminating similar duplicate records are the most important content in the field of data cleaning [16, 12]. When judging whether two records are similar or repeated, they are commonly based on field matching and record matching. At present, more commonly used field matching algorithms are: basic field matching algorithm, recursive field matching algorithm, Smith Waterman algorithm and R-S-W algorithm [17]. Basic nearest neighbor sorting algorithm is commonly used, including several improved algorithms such as multi neighbor sorting method and priority queue algorithm. There are many research achievements made in the elimination of approximate duplicate records, such as clustering analysis, basic clustering tree and priority queue strategy [18].

2.1.3. Data integration

Data integration aims to map the structure and data to the target structure and domain [14]. In recent years, some mature theoretical framework models have emerged, such as AJAX (Asynchronous Javascript and XML) [19], Trillium[20] and Bohn[21]. At the same time, the major database vendors also provide some basic data cleaning tools [22] to fix errors by writing scripts or using data extraction, transformation and loading tools to eliminate inconsistencies. To further strengthen the scalability of these data cleaning tools, some researchers have proposed corresponding models and languages [23] based on the framework of data cleaning systems, such as Merge, Cluste and so on.

2.2. Data mining algorithm

International Conference on Data Mining selected ten classical algorithms of data mining in December 2006. They are K-Means algorithm, SVM algorithm, NBC algorithm, Apriori algorithm, C4.5 algorithm, EM (Expectation Maximization) algorithm, AdaBoost algorithm, kNN (k-Nearest Neighbor) algorithm, PageRank algorithm, and CART algorithm. But in electric power engineering, K-Means algorithm, SVM algorithm, NBC algorithm, ANN algorithm and Decision Tree algorithm are widely used.

2.2.1. Data mining algorithms used in electric power engineering

(1) ANN algorithm

ANN algorithm is a research hotspot in information science, brain science, neuropsychology and other disciplines [24]. It is an artificial neural network built on the basis of human understanding of the brain neural network. It is also a mathematical model of the theoretical human brain neural network. It has the distinctive features of massively parallel processing, distributed information storage, and good self-organizing and self-learning ability [25].

BP (Back Propagation) neural network algorithm, also known as the error back propagation algorithm, is a supervised learning algorithm [26] in ANN. The BP neural network algorithm can theoretically approximate any function. The basic structure is composed of nonlinear variable elements and has strong nonlinear mapping ability. Moreover, the parameters of the middle layer, the number of processing units of each layer and the learning coefficient of the network can be set according to the specific circumstances. It has a wide application prospect in fields of optimization, signal processing and pattern recognition, intelligent control, fault diagnosis and etc. Although the structure of BP neural network is complex, the training time is long, the result is not easy to understand, it has high acceptance ability for noisy data and high accuracy and is preferable in data mining [27].The advantage of rough set data mining algorithm is parallel execution and description of uncertain and incomplete information, as well as the rapid processing of redundant data. However, the problem of rough set is that it is sensitive to object noise. The data mining algorithm combined rough set theory and BP neural network can give full play to their respective advantages, which can overcome the influence of rough set on noise data sensitivity, reduce the training time of BP neural network and provide network convergence [28].

The structure of BP neural network, as shown in Fig. 2, assumes that the output of the hidden node is y_i, the output of the output node is O_k, the expected output of the output node is t_k, and the acti-

vation function of the neuron node is S type function.

$$f(x) = \frac{1}{1 + e^{-x}} \tag{1}$$

The error formula and procedure for each sample of BP neural network are as follows:

① The output of the hidden node:

$$y_i^l = f\left(\sum_j w_{ij} x_j^{-1} - \theta_i\right) = f\left(\text{net}_i^l\right) \tag{2}$$

② Output of the output node:

$$O_k = f\left(\sum_j w_{Mi} y_i^{M-1} - \theta_k\right) = f\left(\text{net}_k\right) \tag{3}$$

③ The error of the output node:

$$E = \frac{1}{2}\sum_k (t_k - O_k)^2 \tag{4}$$

④ Output node mathematical expression:

$$\frac{\partial E}{\partial w_{Mi}} = \sum_{k-1}^{n} \frac{\partial E}{\partial O_k}\frac{\partial O_k}{\partial w_{Mi}} \tag{5}$$

⑤ The derivation of hidden nodes:

$$\frac{\partial E}{\partial w_{ij}} = \sum_k \sum_i \frac{\partial E}{\partial O_k}\frac{\partial O_k}{\partial y_i^{M-1}}\frac{\partial y_i^{M-1}}{\partial w_{ij}} \tag{6}$$

The number of training samples is related to the accuracy of neural network, and the root mean square error function can reflect the learning performance quantitatively. The root mean square function is defined as:

$$e = \sqrt{\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\left(d_{ij} - y_{ij}\right)}{m \cdot n}} \tag{7}$$

In the formula, m refers to the number of training samples and n means the output unit of the neural network.

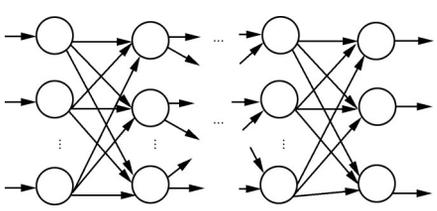

Fig. 2 Structure of BP neural network.

The data mining algorithm flow is based on rough set theory and BP neural network is shown in Fig. 3: step 1: extracting data, clarifying the conditions of mining and determine the mining target; step 2: removing redundant attributes based on rough set theory; step 3: the attribute specification is carried out by the rough set theory; step 4: cost calculation, comparing training data with protocol data; step 5: designing neural network based on training data samples, and using these training data samples to train; step 6: outputting the final result.

(2) Decision tree algorithm

Decision tree algorithm is a method to approximate the value of discrete function, and it is a typical classification method. According to the induction algorithm of data, readable rules and decision trees are generated, and then, new data is handled by decision trees. In essence, this algorithm is the process of classifying data through a series of rules [29,30].

How to construct high precision and small scale decision tree is the core of decision tree algorithm. The decision tree structure can be divided into two steps: The first step is to generate the decision tree from the training sample set. In general, training sample data sets are used for data analysis and processing according to actual needs and a certain degree of integration; the second step is to prune the decision tree, meaning the process of checking, correcting and repairing the decision tree generated in the last stage. It mainly uses the new sample data set (called the test data set) to verify the preliminary rules produced in the decision tree generation process, and cut off the branches that affect the balance accuracy.

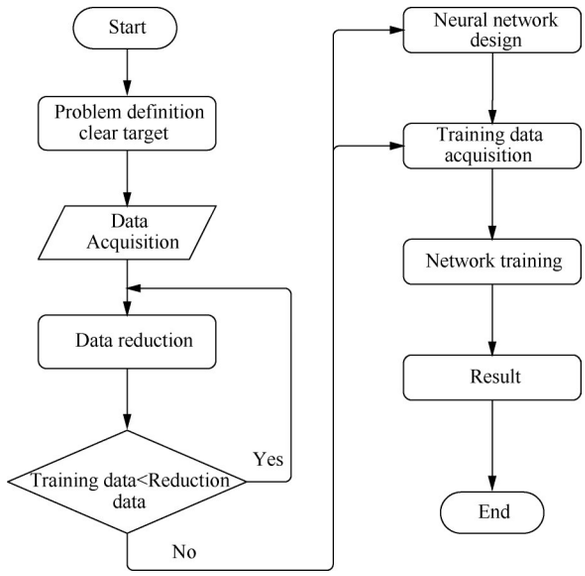

Fig. 3 Flow chart of algorithm based on rough set theory and BP neural network.

ID3 is an algorithm used to construct a decision tree, which takes the decline speed of information entropy as the criterion for selecting test attributes. Each node selects an attribute with the highest information gain that has not yet been used as a partition standard, and then continues the process until the generated decision tree can perfectly classify the training sample [30].

A decision tree can estimate the correct category for a sample E, the amount of information required is:

$$I(p,n) = -\frac{p}{p+n}\log_2^{\frac{p}{p+n}} - \frac{p}{p+n}\log_2^{\frac{n}{p+n}} \tag{8}$$

In the formula, p and n represent the number of positive simple and counter simple sets of E respectively, taking attribute A as the root of decision tree, then the information entropy after classification is

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I\left(p_i, n_i\right) \tag{9}$$

In the formula, $I(P_i,n_i)$ represents the information entropy of subset E_i, and the information content of each subset E_i can be expressed as

$$I(E_i) = \sum_{j=1}^{C} -\frac{P_{ij}}{\left|E_i\right|} * \log\frac{P_{ij}}{\left|E_j\right|} \tag{10}$$

Based on this, the entropy of A is determined:

$$E(A) = \sum_{i=1}^{v} \frac{\left|E_i\right|}{\left|E\right|} * I\left(E_i\right) \tag{11}$$

Select attribute A^* to minimize $E(a)$ and maximize information gain.

The algorithm flow of the decision tree is shown in Fig. 4. Step 1: preprocessing to obtain the sample set and attribute list; step 2: determining whether the attribute is empty, if empty, returning an empty tree, or creating a hollow tree; step 3: calculating the conditional entropy of each attribute, and using the conditional entropy of the minimum value as best attribute; step 4: looping the next cycle for each value in best attribute and return to the tree after the

loop ends; step 5: determining whether the information entropy is zero or not, if so, creating a single node tree and adding it to the tree. Otherwise, the new attribute and new sample are extracted, completing the recursive procedure and return the recursive result to a tree.

(3) SVM algorithm

SVM algorithm is a common method of discrimination [31-33]. The principle of that is to map vectors into a higher dimensional space, and establish a maximum interval hyperplane in this space, which maximizes the distance between two parallel hyperplanes. These two planes are built on both sides of the hyperplane. The larger the distance between the parallel hyperplanes is, the smaller the total error of the classifier will be.

When the classification problem cannot be solved by linear classification in the low latitude space, the data of the low latitude space can be mapped to the high latitude characteristic space to achieve the purpose of linear separable, as shown in Fig. 5.

The key to the transformation from low latitudes to high latitudes lies in finding a function. Which is:

$$\begin{aligned}\left\langle \phi(x_1),\phi(x_2)\right\rangle &= \left\langle (x_1^2,\sqrt{2}\,x_1y_1,y_1^2),(x_2^2,\sqrt{2}\,x_2y_2,y_2^2)\right\rangle \\ &= x_1^2x_2^2 + 2x_1y_1x_2y_2 + y_1^2y_2^2 \\ &= (x_1x_2 + y_1y_2)^2 \\ &= \left\langle x_1,x_2\right\rangle^2 \\ &= K(x_1,x_2)\end{aligned} \tag{12}$$

In order to solve the linear and inseparable problem of mapping from kernel function to high-dimensional space we need relaxation variables:

$$\begin{cases}\min \dfrac{1}{2}\sqrt{\|w\|^2} \\ y_i[(w\cdot x_i)+b] \geqslant 1-\xi(i=1,2,\cdots,n) \\ \xi \geqslant 0\end{cases} \tag{13}$$

The algorithm flow of SVM is shown in Fig. 6. Step 1: data preprocessing; step 2: the SVN optimization equation is constructed and solved, and the interface parameters are obtained; step 3: calculating the classification values from the interface parameters; step 4: classifying the data into second categories according to the clas-

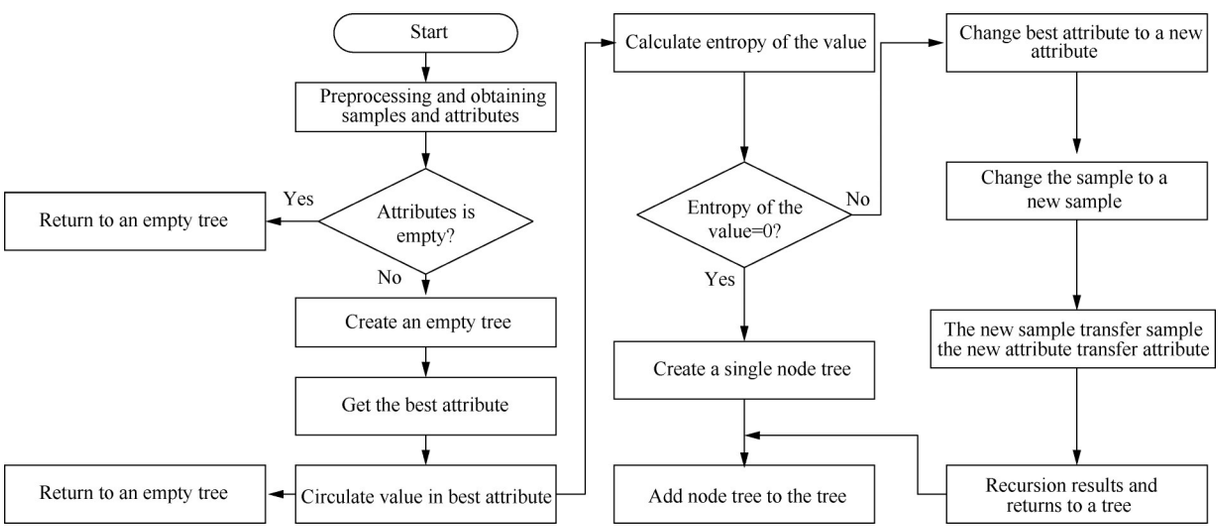

Fig. 4 Flow chart of decision tree.

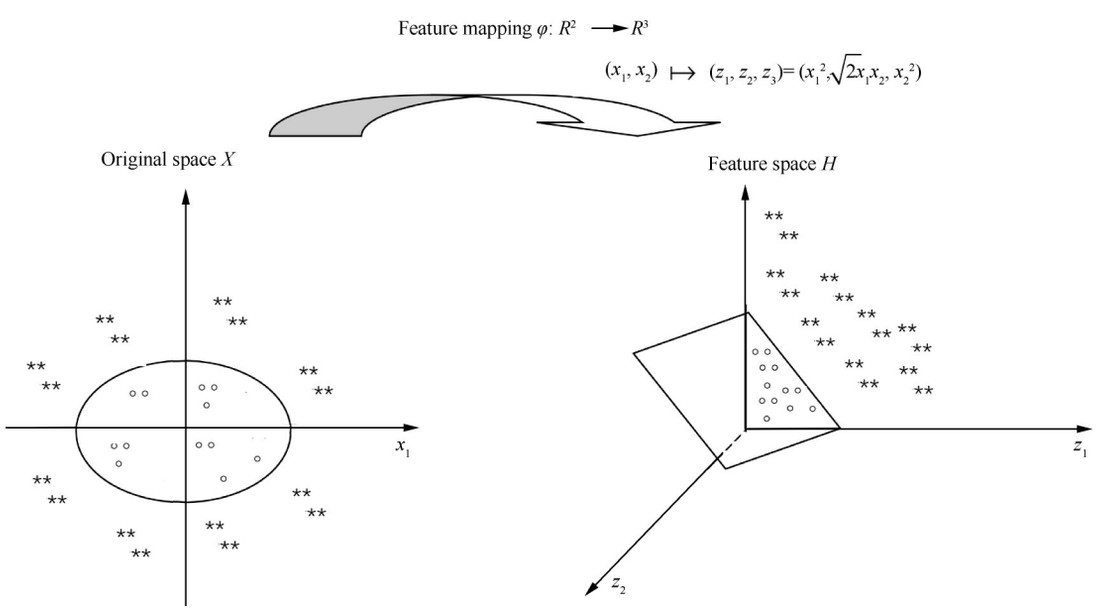

Fig. 5 Mapping from low latitudes to high latitudes.

sification values.

SVM algorithm is a supervised learning method, which is widely used in statistical classification and regression analysis, and it also has advantages of superior generalization performance, global convergence and insensitive sample dimension, and can better solve the problem of nonlinear, high-dimension, and local extreme small point. SVM algorithm shows strong recognition performance in computer aided detection, 3D target recognition, text recognition, face recognition, protein prediction, facial expression classification and speech recognition [34,35].

2.2.2. Other data mining algorithms

(1) Apriori algorithms

Apriori algorithm is the most influential algorithm for mining frequent item sets of Boolean association rules [36]. Its core is to recursive algorithm based on the idea of two stage frequency set. The frequent itemset is extracted from the two stages through candidate set generation and closing down detection of the plot. The association rules belong to single dimension, monolayer and Boolean association rules in classification. Apriori algorithm is a frequent itemset algorithm for mining association rules, and the algorithm has been widely applied to business, network security and other fields.

(2) NBC algorithm

NBC algorithm is one of the most widely used classification models [37] which have a solid mathematical foundation and a stable classification efficiency. The estimated parameters are very few and are not sensitive to missing data, and the algorithm is simple and the error is small. NBC model assumes that the attributes are independent of each other. This assumption is not often valid in practical applications, which has had certain impact on the correct classification of NBC models. When the number of attributes is large or the correlation between attributes is large, the classification efficiency of the NBC model is not as good as that of decision tree model. However, when the correlation between attributes is small, the NBC model performs best.

(3) K-means algorithm

K-means algorithm is a partition based the widely used clustering algorithm [38]. The objects of n are divided into k partitions according to their attributes. Assuming that the attribute of the object comes from the space vector, the goal is to minimize sum of the mean square error within each group [39]. From the performance of the algorithm, it does not guarantee that the global optimal solution will be obtained. The quality of the final solution depends to a large extent on the initialized grouping. This algorithm is very fast, which can be computed many times to select the optimal solution. Moreover, algorithm is simple, implementable and extensible, which can also handle large data sets.

The characteristics and complexity of these algorithms as shown in Table 1.

2.3. Data mining software

With the development of data mining applications, data mining software is closely combined with the following three aspects: ① database and data warehouse; ② multiple types of data mining algorithms; ③ data cleaning, conversion and other preprocessing work. Initial stage of data mining software is independent data mining software. However, the problems in the real world are varied,

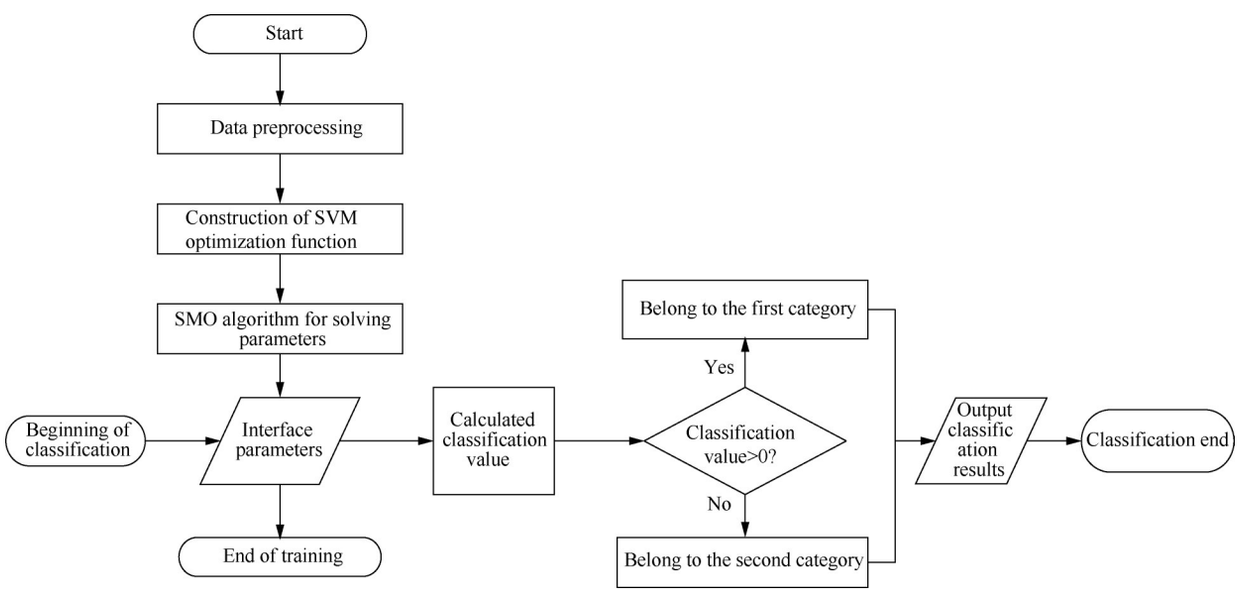

Fig. 6 Flow chart of SVM algorithm.

Table 1 Characteristics and complexity of algorithms.

Algorithm	Characteristics		Complexity	
	Whether iterative	Limitation	Time complexity	Space complexity
ANN	Yes	May fall into local extremes	$O(m^2)$	$O(m)$
K-means	Yes	K value is random	$O(m)$	$O(m)$
Decision tree	Yes	Over fitting problems	$O(m^2)$	$O(m)$
SVM	Yes	Suitable for small sample, large data sensitive	$O(m)$ (m^2)	$O(m)$
Apriori	No	Control of frequent itemsets	Exponential	Exponential
NBC	No	Dimension sensitive, large data sensitive	$O(dm)$	$O(m)$

which makes data mining software go through second stages: horizontal data mining tool set. With the increasingly application of horizontal data mining tools, it is found that only those experts who are proficient in data mining algorithms can skillfully use these tools. Since then, a large number of data mining tools have begun to provide a vertical data mining solution, which also makes data mining software experience third stages: a vertical data mining solution. The fourth generation software can excavate various types of data produced by embedded systems, mobile systems, and ubiquitous computing devices. At present, mobile computing is becoming more and more important. The combination of data mining and mobile computing is the current research hotspot. The main characteristics of each stage of the product can be summarized as shown in Table 2.

3. Application of data mining technology in electric power engineering

Data mining technology is widely applied in all walks of life, including finance, health care, marketing, retailing, manufacturing, justice and insurance. The scientific and technological innovation made based on data mining technology is well known in the news magazine and other media. In order to make the power engineering researchers further understand the advance of technology, data mining technology in electric power engineering is introduced in details.

In this engineering, digital technology has been widely applied in recent years [40]. Data acquisition and monitoring system, distribution network management system, power quality monitoring system, as well as some auxiliary decision-making systems, such as power metering system and quotation processing system can collect and record the operation information of the power grid in real time. The application of these systems has produced massive data. So how to make full use of the data and analyze, process, extract and excavate useful knowledge quickly and effectively have become an urgent problem to be solved in the power industry.

3.1. Prediction

3.1.1. Electricity price prediction

Electricity price is the most effective method of power resource allocation. Accurate price forecasting can provide investment guidance for market participants and further avoid risks. In order to improve the precision of electricity price prediction, Mori [41] established a hybrid prediction model based on ANN and SVM in data mining and traditional time series method. This method can predict more accurately. In order to better analyze the fluctuation of electricity price, Zhao [42] first proposed the concept of price spike. Time and amplitude prediction model of the price spike was established to help participants in the electricity market to better analyze the possible abnormal electricity price. Lu[43] proposed a data mining based electricity price forecast framework, which can predict the normal price as well as the price spikes. He also explored the reasons of price spikes based on the measurement of a proposed composite supply-demand balance index and relative demand index, and the model was able to generate forecasted price spike, level of spike and associated forecast confidence level. As for medium to long-term price forecasting, Florian Ziel [44] introduced a new approach to simulate electricity prices with hourly resolution from several months up to three years. Considering the uncertainty of future events, they were able to provide probabilistic forecasts which can detect probabilities of price spikes even in the long-run. Patil [45] attempted to use K-means and k-NN algorithm to divide historical prices data instead of calendar, and use ARIMA (Auto Regressive Integral Moving Average) statistical model to predict the short-term price of electricity, which proves the validity of the model. Wu[46] used K-means algorithm to classify of data of historical prices of New York Energy Market and the performance of forecasting model is very satisfactory.

3.1.2. Power load forecasting

Accurate load forecasting is the key for the power dispatching department to formulate the power supply plan and realize the balance between supply and demand. Lambent-Tomes [47] applied data mining technology to power load forecasting, which used fuzzy set theory and fuzzy approximation theory, and the short-term prediction target of one year has been realized. It has been successfully applied to Brazil power grid for many years. Yuniarti[48] proposed a method using data mining technique for forecasting electrical load. The method of combining wavelet transform with packet processing is applied to short-term power load forecasting. The results show that the method has good accuracy and improves the forecasting of electricity on average above 50% per year. Wang [49] analyzed

Table 2 Development of data mining software.

Generation	Representative	Features	Data mining algorithm	Integration	Distributed computing model	Data model
First	Early CART System in Salford Systems Company	An independent application	Support one or more algorithms	Independent system	Single machine	Vector
Second	DBMiner SAS Enterprise Miner	Integrated with databases or data warehouses	Multiple algorithms; capable of mining data that cannot be placed into memory at one time.	Data management system, including database and data warehouse.	A group of computers in a local area	System supports objects, text and continuous media data.
Third	SPSS Clementine	Predictable model, system integration	Multiple algorithms	Data management and prediction model system	Internet/extranet network computing	Support semi-structured data and web data
Fourth	IBM DB2 Intelligent Miner Scoring Service	Data combination with mobile data / various computing devices	Multiple algorithms	Data management, prediction model, mobile system	Mobile and various computing devices	Universal computing model

deeply regional electricity load through data mining technology based on variation function theory and its structural analysis, and the author realized the unbiased optimal estimation on regional variations in finite regions through spatial local difference algorithm. A standardized algorithm of electric power data based on Kriging interpolation method was put forward. Xydas [50] forecasted electric vehicle load by algorithm of decision tables, decision trees, ANN and SVM, the results showed that data mining methods can be used for forecasting the EV charging load, with increased accuracy especially when the configuration parameters of each method are carefully selected. As for short-term power load forecasting, SVM algorithm which is based on data mining is very effective and achieves higher prediction accuracy [51].

3.2. Classification

3.2.1. Fault type identification

The fault types of power system are usually divided into single-phase to ground short circuit, two phase short circuit and three phase short circuit. The fault mode space is generally nonlinear, because the voltage and current signal are affected by the operating mode of the system, the fault location and the cause of the fault. Many scholars use data mining technology to classify power system faults. To solve the problem of relay response, the optimal NBC algorithm is utilized by Faiz [52] to develop a method for discriminating the fault from non-fault events, and the proposed method has been designed based on extracting the modal parameters of the current waveform using the Prony method. Babnik [53] used NBC method to classify transformer faults into internal or external grounding and short-circuit faults. The results of all methods show that they can identify power transformer faults quickly and successfully. As for fault locating in a radial power distribution system, Kumar [54] used database to act as the trainer to the fuzzy expert system, and the proposed method aims to lighten the decision-making burden on some of the system operators. Yan [55] adopted the time series model in data mining to analyze the running data in the process enterprise, a sequential association rule model is obtained to arrange the influence of abnormal parameter points in time order on equipment failure, which plays a role in the warning and monitoring of equipment fault. In Zhang's [56] research, global information was introduced into the electric power system, and he mainly used cluster analysis technology of data mining theory to resolve quickly and exactly detection of fault components and fault sections, and finally accomplished fault analysis. Xu [57] proposed PDFCC (power distribution fault cause classifier) based on data mining classification method to address the identification problem caused by fault in the power distribution systems.

3.2.2. System state classification

Each monitoring system in the power system can acquire real-time parameters such as power generation level, power flow distribution, load level and fault condition. These parameters can effectively respond to the state of the power system, which is conducive to the reasonable dispatching of the power grid operators and maintaining the safe and stable operation of the system. Data mining technology can classify the state of the system so as to facilitate management and monitoring. In order to improve the efficiency of the system, Lambert-Tomes [58] constructed classification rules based on rough set theory to remove redundant data and obtained effective data, and divided the system into normal, abnormal and restorative states. Huang [59] determined whether the system is in a normal state through a decision tree or an extreme contingency to determine whether protection is necessary. The wide application of the distributed power supply made it necessary for the operators to judge whether the isolated island state has appeared in the systems with multiple power sources, and the data mining was introduced into the island detection by El-Arroudi [60], and the method uses and combines the parameters of various system parameters to ensure the security of isolated island detection. The effectiveness of the algorithm is verified by experiments and the application scope of data mining is further expanded.

3.3. Analysis of association rules

There are many correlations among the variables in the power system, such as the correlation between the electricity consumption and the electricity price, the precipitation and the temperature, the correlation between the voltage and the harmonic current at different locations, the correlation between the coal consumption of the generator and the main steam pressure, the main steam temperature, the water supply temperature, and other controllable parameters. Data mining technology can be used to analyze the association rules in the power system, which can provide guidance for improving the efficiency of power generation, thus optimizing the power transmission and reducing the cost of facilities.

3.3.1. Generation side association rules

In view of the research of photovoltaic array generation forecasting method, Cheng [61] proposed a method for forecasting photovoltaic power generation by using forward selection and K-means clustering and radial basis function neural network. The experimental results showed that compared with the traditional neural network prediction model, the model has fewer input variables and higher prediction accuracy. For thermal power generation, data mining technology can also improve the efficiency of power generation. In view of the characteristics of the numerical operation parameters of the thermal power unit, Niu [62] proposed a fuzzy association rule data mining method based on the improved Apriori algorithm and established the association rules between the parameters of the boiler exhaust temperature and the efficiency of the boiler, and obtained the learning rules to improve the economy of the unit. Cui [63] pointed out that the cost of increasing coal-fired boilers caused by flue gas denitration is the cost of boiler efficiency loss, and the mathematical formula was given by him, but the influence of the fluctuation of power coal price on the cost of coal-fired boilers has not been considered. Cai [64] used data mining to take advantage of inherent tolerance of the imprecision and uncertainty to obtain tractability, robustness, and low solution-cost. As for forecasting the generation of solar power, Khabibrakhmanov [65] forecasted solar power generation by using real-time power data, weather data, and complexity-based similarity factors, which was achieved by data mining technology, and he has applied for the patent protection.

3.3.2. Analysis of data association of power grid operation

The data mining technology can locate the association rules between the fault phenomenon or the fault cause and category, and find correlation characteristics between fault elements of the power grid, which can be put forward as a strategy to monitor and diagnose fault equipment [66]. Li[67] proposed a method of fault diagnosis for power grid based on feature data mining. It combined data mining methods such as association rule analysis, sequence pattern analysis, cluster analysis and other expert systems, and gave the data mining algorithm flow. Tong[68] applied artificial intelligence and data mining technology to the transient stability evaluation of power system, analyzed and compared the research results in the aspects of principal component analysis, genetic algorithm, rough set, information entropy preprocessing, ANN and SVM classifier and

visual display. Zeng[69] introduced the application of data mining technology in solving the problem of load optimal allocation in real-time factory level, and used multi factor weight allocation method based on information entropy to establish a real-time discrete model between load and power supply coal consumption rate. Xu[70] introduced the big data technology for power transmission line fault analysis, in which Apriori algorithm is used to mine the key attribute of fault cause and fault parts of the transmission line. The results showed that the proposed scheme can effectively select the key attributes of cause and location of the power transmission line fault. SSAE (stacked sparse autoencoder) is more effective in solving power system fault diagnosis due to its network structure and layer-wise training mechanism. Wang[71] proposed SSAE-based network with SVM to improve the accuracy of fault diagnosis in power systems, which has achieved ideal effect.

3.4. Clustering and outlier analysis

Unlike classification, clustering is to divide data into multiple classes or clusters, so that objects in the same cluster more similar and the objects in the different clusters are less similar. Data mining technology is used to cluster different power users or generators to obtain different class attributes. Chicco[72] described the electrical load changes caused by anomalous days (holidays, working days between holidays, social events) by using the methods of Kohonen map with a classic clustering algorithm and an ANN-based approach, and the comparison is then made. The results showed that the combination of use of both clustering techniques allows better understanding of the anomalous load patterns. Mori [73] proposed an efficient ANN method to forecast electricity price, and a clustering technique is used to determine the center of RBFN (radial basis function network) and NRBFN (normalized radial basis function network). The effectiveness of the proposed method is demonstrated for real data of hourly electricity price for ISO New England. Liu [74] adapted K-means clustering algorithm to analyze customer load and similar behavior between electricity users, and the method of principal component analysis was used on the clustering results visualization, fully proving the rationality and correctness of the clustering. Damayanti [75] compared three methods of clustering techniques, namely the K-means, fuzzy and C-means, and found that K-means was the most appropriate method to classify the electrical load profile.

In the process of data analysis, such data objects are often found because they are significantly different from other data and are called Outliers. Outliers mean that there are abnormal situations such as faults. Outlier analysis and detection is also a significant task of data mining. Dessertaine[76] used ANN method to correct load outlier data as a preprocessing step of load forecasting. Neagu[77] adapted a statistical based data mining for load curves characterization by detecting outliers` information provided by Smart Meters in real distribution networks. After eliminating outliers, the remaining data had led to the discovery of accurate patterns that are characterized very well to the load curves characteristics through indicators. Stealing and leakage of electricity has always been challenging power supply enterprises. Tang[78] studied of the abnormal electricity consumption detection system based on the outlier behavior pattern recognition and provided a reference for the researchers of anti-power-stealing. Sun[79] focused on the outlier detection of electricity consumption data, and introduced the causes of electricity consumption outlier data from the negative and positive aspects respectively. Moreover, he also provided a review on the detection methods of electricity consumption outlier data on the basis of data mining. As a data cleaning process, outlier analysis is also applied in improving power system state assessment[80] and Data Debugging accuracy[81].

4. Conclusions

4.1. Summary

Because of the rapid development of computer industry, data mining technology has almost become mature in theory, especially in algorithms, dozens of which can achieve various functions and provide support for data processing in electric power engineering. Data mining technology has valuable information in massive data, makes optimization strategy, improves efficiency, reduces costs and promotes the development of electric power engineering.

Data mining technology is widely used in the field of power engineering. In price and load forecasting, the data mining algorithm of ANN, SVM, K-means and decision trees can improve the precision of the prediction. In addition, these methods are also commonly used in power generation side association rules, power grid operation data association analysis and clustering and outlier analysis; NBC, decision trees and cluster analysis are commonly used in power failure, system state and other classification.

ANN, SVM and decision tree algorithm are mainly used in building models, realizing prediction and clustering correlation functions. K-means can also implement model prediction and association functions. NBC algorithm has been applied widely in fault classification. Apriori algorithm has unique advantages in establishing associated physical quantities.

4.2. Expectation

Data mining technology still faces many problems and challenges, such as the efficiency of data mining in the ultra large data set, the development of mining methods adapted to multiple data types, noise tolerance, data mining in the network and distributed environment, dynamic data mining and data mining. The followings are the important future trends of data mining:

The standardized description of data mining language: the standard data mining language will help the systematic development of data mining, improve the interoperability between multiple data mining systems and functions and promote their use in enterprises and society.

Visualized data mining process can seek the visualization method in the process of data mining, making it is easy to understand and manipulate the process of knowledge discovery for the users, which can make the data mining process become a part of the user's business process, and also facilitate the human interaction in the process of knowledge discovery.

The mining technology combines various heterogeneous data to exploit various unstructured data (Data Mining for AMD), such as the exploitation of text data, graphic data, video image data, sound data and even comprehensive multimedia data.

The authors declare that there is no conflict of interest regarding the publication of this article.

References

1. Naisbitt J. Megatrends: Ten new directions transforming our lives. *Business Horizons* 1983; 26(3):84-6.
2. Søilen K S. An overview of articles on competitive intelligence in JCIM and CIR. *Journal of Intelligence Studies in Business* 2013; 3(1):44-58.
3. Haixia G. New social networking features and model analysis of information dissemination. *Journal of Modern Information* 2012; 1: 56-9.

4. Uthurusamy R. From data mining to knowledge discovery: Current challenges and future directions. *Advances in knowledge discovery and Data Mining.* 1996.p.561-9.
5. Klösgen W. Knowledge discovery in databases and data mining. *International Symposium on Methodologies for Intelligent Systems*. 1996. p.623-32.
6. Feyyad U M. Data mining and knowledge discovery: Making sense out of data. *IEEE Expert* 1996; 11(5): 20-5.
7. Frankish K, Ramsey WM. *The Cambridge handbook of artificial intelligence*. Cambridge：Cambridge University Press. 2014.
8. Yilong G. Data Mining and Its application in Engineering diagnosis [dissertation]. Xi'an: Xi'an Jiaotong University. 2000.
9. Gaber M M, Zaslavsky A, Krishnaswamy S. Mining data streams: a review. *ACM Sigmod Record* 2005; 34(2): 18-26.
10. Jiang N, Gruenwald L. Research issues in data stream association rule mining. *ACM Sigmod Record* 2006; 35(1): 14-9.
11. Gray J, Chaudhuri S, Bosworth A, et al. Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals. *Data Mining and Knowledge Discovery* 1997; 1(1): 29-53.
12. Florescuand D. An extensible framework for data cleaning. *Proceedings of the 16th International Conference on Data Engineering.* 2000. p.312.
13. Woodard M, Wisely M, Sarvestani S S. A survey of data cleansing techniques for cyber-physical critical infrastructure systems. *Advances in Computers. Elsevier* 2016; 102: 63-110.
14. Zolhavarieh S, Aghabozorgi S, Teh Y W. A review of subsequence time series clustering. *The Scientific World Journal* 2014; 2014:312521.
15. Kaur DP, Walia AS. A study on clustering based methods. *International Journal of Advanced Research in Computer Science* 2017; 8(4).
16. Hernández MA, Stolfo SJ. Real-world data is dirty: Data cleansing and the merge/purge problem. *Data Mining and Knowledge Discovery* 1998; 2(1): 9-37.
17. Monge AE, Elkan C. The field matching problem: algorithms and applications. *Proc Acm International Conference on Knowledge Discovery & Data Mining.* 1996.p.267-70.
18. Hu W, Zaveri A, Qiu H, et al. Cleaning by clustering: methodology for addressing data quality issues in biomedical metadata. *BMC Bioinformatics* 2017; 18(1): 1-12.
19. Galhardas H. Data cleaning and transformation using the AJAX framework. *International Summer School on Generative and Transformational Techniques in Software Engineering*. 2005.p.327-43.
20. Harte-Hanks Trillium Software. [2007-01-09]. http://www.trilliumsoftware.com.
21. Bruckner RM, List B, Schiefer J. Striving towards near real-time data integration for data warehouses. *International Conference on Data Warehousing and Knowledge Discovery*. 2002.p.317-26.
22. Devi S, Kalia A. Study of data cleaning & comparison of data cleaning tools. *International Journal of Computer Science and Mobile Computing* 2015; 4(3): 360-70.
23. Galhardas H, Florescu D, Shasha D, et al. Declarative data cleaning: Language, model, and algorithms. Report No. RR-4149. INRIA, 2001.
24. Kamruzzaman SM, Sarkar AM. A new data mining scheme using artificial neural networks. *Sensors* 2011; 11(5): 4622-47.
25. Sinkov A, Asyaev G, Mursalimov A, et al. Neural networks in data mining. *2016 2nd International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM)*. 2016.p.1-5.
26. Dan SU. Research on high-altitude meteorological data mining method based on BP neural network. *Modern Electronics Technique* 2017; 40 (24):40-2.
27. Zhang D, Jiang Q, Li X. Application of neural networks in financial data mining. *International Conference on Computational Intelligence.* 2004.p.392-5.
28. Si L, Liu X, Tan C, et al. A novel classification approach through integration of rough sets and back-propagation neural network. *Journal of Applied Mathematics* 2014;(3):1-11.
29. Jin M, Wang H, Zhang Q, et al. Financial management and decision based on decision tree algorithm. *Wireless Personal Communications* 2018; 102(4): 2869-84.
30. Sudrajat R, Irianingsih I, Krisnawan D. Analysis of data mining classification by comparison of C4. 5 and ID algorithms. *Materials Science and Engineering Series* 2017; 166(1): 012031.
31. Veale M, Brass I. Administration by algorithm? Public management meets public sector machine learning. *Public Management Meets Public Sector Machine Learning.* 2019.
32. Cortes C, Vapnik V. Support-vector networks. *Machine Learning* 1995; 20(3): 273-97.
33. Vapnik V, Golowich SE, Smola A. Support vector method for function approximation, regression estimation, and signal processing. *Advances in Neural Information Processing Systems.* 1997.p.281-7.
34. Huang CL, Chen MC, Wang CJ. Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications* 2007; 33(4): 847-56.
35. Song J, Tang H. Support vector machines for classification of homo-oligomeric proteins by incorporating subsequence distributions. *Journal of Molecular Structure*: *THEOCHEM* 2005; 722(1-3): 97-101.
36. Agrawal R, Srikant R. Fast algorithms for mining association rules. *Proceedings of the 20th VLDB Conference*. 1994.p.487-99.
37. Rennie JD, Shih L, Teevan J, et al. Tackling the poor assumptions of naive bayes text classifiers. *Proceedings of the 20th international conference on machine learning* (ICML-03). 2003.p.616-23.
38. Han J, Kamber M, Pei J. Data mining concepts and techniques third edition. *The Morgan Kaufmann Series in Data Management Systems* 2011;5(4): 83-124.
39. Agrawal R, Gehrke J, Gunopulos D, et al. Automatic subspace clustering of high dimensional data for data mining applications. *Proceedings of the 1998 ACM SIGMOD international conference on Management of data.* 1998.p.94-105.
40. Fitzgerald M, Kruschwitz N, Bonnet D, et al. Embracing digital technology: A new strategic imperative. *MIT Sloan Management Review* 2014; 55(2): 1.
41. Mori H, Awata A. A hybrid method of clipping and artificial neural network for electricity price zone forecasting. *2006 International Conference on Probabilistic Methods Applied to Power Systems*. 2006.p.1-6.
42. Zhao JH, Dong ZY, Li X, et al. A framework for electricity price spike analysis with advanced data mining methods. *IEEE Transactions on Power Systems* 2007; 22(1): 376-85.
43. Lu X, Dong ZY, Li X. Electricity market price spike forecast with data mining techniques. *Electric Power Systems Research* 2005; 73(1): 19-29.
44. Ziel F, Steinert R. Probabilistic mid-and long-term electricity price forecasting. *Renewable and Sustainable Energy Reviews* 2018; 94: 251-66.
45. Patil M, Deshmukh SR, Agrawal R. Electric power price forecasting using data mining techniques. *International Conference on Data Management*, *Analytics and Innovation* (*ICDMAI*). 2017.p.217-23.
46. Wu X, Zhou H. Short-term electricity price forecasting based on subtractive clustering and adaptive neuro-fuzzy inference system. *Power System Technology* 2007; 31(19): 69-73.
47. Lambert-Torres G, Marra W, Lage WF, et al. Data mining in load forecasting: an approach using fuzzy techniques. *2006 IEEE Power Engineering Society General Meeting*. 2006.
48. Yuniarti T, Surjandari I, Muslim E, et al. Data mining approach for short term load forecasting by combining wavelet transform and group method of data handling (WGMDH). *2017 3rd International Conference on Science in Information Technology* (ICSITech). 2017.p.53-8.
49. Wang Q, Sun Q, Li Q, et al. A method of electricity utilization load analysis and visualization based on data mining of electric power big data. *DEStech Transactions on Engineering and Technology Research.* 2016.
50. Xydas S, Marmaras CE, Cipcigan LM, et al. Electric vehicle load forecasting using data mining methods. *Hybrid and Electric Vehicles Conference.* 2014.p.1-6.

51. Sun F, Yang Y. *A research on power load forecasting model based on data mining. research and practical issues of enterprise information systems II*. 2008.p.1369-77.
52. Faiz J, Lotfi-fard S, Shahri SH. Prony-based optimal bayes fault classification of overcurrent protection. *IEEE Transactions on Power Delivery* 2007; 22(3): 1326-34.
53. Babnik T, Gubina F. Fast power transformer fault classification methods based on protection signals. *IEE Proceedings-Generation, Transmission and Distribution* 2003; 150(2): 205-10.
54. Kumar N, Sharma M, Sinha A, et al. Fault detection on radial power distribution systems using fuzzy logic. *International Journal of Electrical and Electronics Engineers* 2015; 398-406.
55. Yan W, Zhang H, Lu JF. Study and application of tin e- interval sequential pattern to equipment fault monitoring. *Journal of Computer Applications* 2005; 25(7):1584-6.
56. Zhang Y, Jing MA, Zhang J, et al. Applications of data mining theory in electrical engineering. *Engineering* 2013; 1(3): 211-5.
57. Xu L, Chow M, A classification approach for power distribution systems fault cause identification. *IEEE Transactions on Power Systems* 2006; 21(1): 53-60.
58. Lambert-Torres G. Application of rough sets in power system control center data mining. *Power Engineering Society Winter Meeting.* 2002. p.627-31.
59. Huang JA, Vanier G, Valette A, et al. Application of data mining techniques for automat settings in emergency control at Hydro-Qudbec. *2003 IEEE Power Engineering Society General Meeting.* 2003.p.2037-44.
60. El-Arroudi K, Joos G, Kamwa I, et al. Intelligent-based approach to islanding detection in distributed generation. *IEEE Transactions on Power Delivery* 2007; 22(2): 828-35.
61. Cheng Z, Li SY, Han LJ, et al. PV power generation forecast based on data mining method. *Acta Energiae Solaris Sinica* 2017; 38(3): 726-33.
62. Niu C, Li J, Liu J, et al. Correlation analysis of operation data and its application in operation optimization in power plant. *2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery.* 2008.p.581-5.
63. Cui C, Liu J, Yang T. Economical optimization of a boiler denitration system based on GA and fuzzy association rules. *Journal of Chinese Society of Power Engineering* 2016; 36(4): 300-6.
64. Cai T. Application of data mining and analysis techniques for renewable energy network design and optimization. *Data Mining and Analysis in the Engineering Field.* 2014.p.33-47.
65. Khabibrakhmanov I, Kumar T, Lavin MA, et al. Forecasting solar power generation using real-time power data, weather data, and complexity-based similarity factors. United States Patent US 9857778. 2018 Jan 2.
66. Lin D, Jun P, Jun T. Fault location for transmission line based on traveling waves using correlation analysis method. *2008 International Conference on High Voltage Engineering and Application.* 2008.p.681-4.
67. Li Z, Bai X, Zhou Z, et al. Method of power grid fault diagnosis based on feature mining. *Proceedings of the CSEE* 2010; 30(10): 16-22.
68. Tong X, Ye S. A survey on application of data mining in transient stability assessment of power system. *Power System Technology* 2009; 33 (20): 88-93.
69. Zeng D, Yang T, Cheng X, et al. Application of data mining method in real-time optimal load dispatching of power plant. *Proceedings of the CSEE.* 2010.p.109-14.
70. Xu P, Xiao F, Feng S, et al. Data mining of power transmission line fault based on Apriori algorithm. *2017 IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference* (ITNEC). 2017.p.49-54.
71. Wang Y, Liu M, Bao Z, et al. Stacked sparse autoencoder with PCA and SVM for data-based line trip fault diagnosis in power systems. *Neural Computing and Applications* 2019; 31(10): 6719-31.
72. Chicco G, Napoli R, Piglione F. Load pattern clustering for short-term load forecasting of anomalous days. *2001 IEEE Porto Power Tech Proceedings*. 2001.
73. Mori H, Awata A. Normalized RBFN with hierarchical deterministic annealing clustering for electricity price forecasting. *2007 IEEE Power Engineering Society General Meeting*. 2007.p.1-7.
74. Liu L. Cluster analysis of electrical behavior. *Journal of Computer and Communications* 2015; 3(5): 88.
75. Damayanti R. Analisis profil beban listrik menggunakan teknik clustering[dissertation]. Indonesia: Universitas Pendidikan Indonesia, 2016.
76. Dessertaine A. Detection of remarkable values in Individual electric consumption's series using non-parametric approach. *2007 IEEE Lausanne Power Tech.* 2007.p.1964-9.
77. Neagu BC, Grigoraș G, Scarlatache F. Outliers discovery from Smart Meters data using a statistical based data mining approach. *2017 10th International Symposium on Advanced Topics in Electrical Engineering* (ATEE). 2017.p.555-8.
78. Tang Y, Wen M, Xu T, et al. The abnormal electricity consumption detection system based on the outlier behavior pattern recognition. *2017 International Conference on Energy, Power and Environmental Engineering*(*ICEPEE2017*). 2017.
79. Sun S, Li G, Chen H, et al. Optimization of support vector regression model based on outlier detection methods for predicting electricity consumption of a public building WSHP system. *Energy and Buildings* 2017; 151: 35-44.
80. Huang SJ, Lin JM. Enhancement of anomalous data mining in power system predicting-aided state estimation. *IEEE Transactions on Power Systems* 2004; 19(1): 610-9.
81. Teeuwsen SP, Erlich I. Neural network based multi-dimensional feature forecasting for bad data detection and feature restoration in power systems. *2006 IEEE Power Engineering Society General Meeting*. 2006.