Research on Data Mining Algorithm for Regional Photovoltaic Generation

Zhen Lei[1] and Yong-biao Yang[2(✉)]

[1] State Grid Jiangsu Electric Power Company, Nanjing, China
[2] Southeast University, Nanjing, China
danghongen2017@163.com

Abstract. Traditional data mining algorithms have problems such as poor applicability, high false positive rate or high false positive rate, resulting in low security and stability of the power system. For this reason, the regional photovoltaic power generation data mining algorithm is studied. Classification of data sources facilitates correlation calculations, and matrix relationships are used to calculate data associations. Combined with the data relevance, the association rules are output, and the output results inherit the clustering processing and time series distribution of the implicit data, thereby realizing the extraction of hidden data and completing the regional photovoltaic power generation data mining. The experimental results show that the regional PV power generation data mining algorithm has high stability and can effectively solve the system security problem.

Keywords: Raw input · Data acquisition · Data mining · Dynamic features

1 Introduction

The regional photovoltaic power generation array is the core component of the photovoltaic power generation system. The solar radiation to the ground energy is directly converted into electrical energy through the photovoltaic effect of the panel. Unlike other conventional energy sources such as thermal power, the output power of photovoltaic power generation systems is greatly affected by weather factors such as solar irradiance, temperature, humidity, wind speed and wind, and has obvious characteristics such as variability, intermittentness and uncertainty. The reliability of power system energy management requires that the generation, distribution and use of electric energy remain in a stable and balanced state for a long time. When large-scale photovoltaic power generation is connected to the large power grid for grid-connected power generation, photovoltaic power generation and grid-connected power generation will generate voltage balance between the two power generation system cabinets, which will have a huge impact on the safe and stable operation of the power system. If you want to make a large-scale photovoltaic power generation system smoothly connect to the large grid and generate electricity, the large power grid needs to provide a certain capacity of rotating standby to offset the fluctuation of photovoltaic power generation. It will cause a lot of waste of resources and reduce the economic benefits of photovoltaic power generation. However, due to energy shortages and environmental

G. Gui and L. Yun (Eds.): ADHIP 2019, LNICST 301, pp. 429–438, 2019.
https://doi.org/10.1007/978-3-030-36402-1_46

problems, it is necessary to develop and utilize solar energy as much as possible. In the face of the problem of grid-connected power generation of large-scale photovoltaic power generation systems, power grid companies are in a dilemma. Therefore, how to play the role of photovoltaic power generation as much as possible under the premise of meeting the safety, stability and economic operation of the power grid has become a problem to be solved in the field of photovoltaic power generation. One of the key technologies to solve this problem is to accurately predict the future PV power output to support the power system's power generation plan, power flow optimization, and adjustment of peak and frequency. Therefore, the topic of this paper has a very important role in the sustainable development and application of photovoltaic power generation [1].

2 Regional Photovoltaic Power Generation Data Mining Algorithm

2.1 Data Source Classification

Existing photovoltaic power generation forecasting methods are mainly divided into two broad categories: direct forecasting and indirect forecasting. The indirect prediction method is carried out in two steps. First, the solar irradiance is predicted by various weather factor prediction data at the solar power station, and then the irradiance is brought into the photoelectric conversion efficiency model to calculate the power generation. The direct prediction rule does not require measurement and prediction of solar irradiance, and is directly modeled based on meteorological information and historical power generation data. Predicting photovoltaic power Compared to indirect forecasting, direct prediction does not require irradiance monitoring and prediction. It is more versatile, less computationally intensive, and more predictive. Therefore, this paper uses the direct prediction method [2].

Based on the historical data of photovoltaic power generation for the past two years and the corresponding weather information, the combination of ambient temperature, humidity, wind speed and AQI is chosen instead of irradiance as the original input. Aiming at the problems of the original input variables and the compatibility, the regression-based forward selection method is used to reduce the original input variables by seasons, and the weather factors with moderate quantity and low correlation are selected as the input variables of the model. Then, for the problem of poor generalization ability of a single model, K-means clustering algorithm is used to cluster historical data, and the data is divided into similar weather types to complete the classification of data sources [3]. Its classification and practical application are as follows:

Table 1 is a classification and practical application. A categorical data source refers to a non-trivial process of extracting previously unknown, regular, and future-ready and understandable knowledge and information from a vast, fragmented, noisy, and fuzzy, random, historical data set. "Before unknown" refers to the fact that data mining information is not recognized by people before, that is, novel information, so the more innovative the information, the greater the value [4]. "Available in the future" means

Table 1. Classification and practical application

Type	Classification	Application example
Standalone photovoltaic Power generation system	Battery-free DC photovoltaic power generation system	DC photovoltaic water pump, charger, solar fan cap
	DC photovoltaic power generation system with battery	Solar flashlight, solar cell phone charger, solar lawn lamp
	AC and AC/DC hybrid photovoltaic power generation system	Microwave relay station and environmental monitoring station of AC solar household system Equipment Small ower stations in power-free areas
	Municipal electricity complementary photovoltaic power generation system	AC solar household system, small power plants in non-electric areas

that the extracted knowledge has potential use value, common-sense knowledge or facts, and unachievable knowledge are meaningless. "Can be understood" requires knowledge of the knowledge or discovered patterns that the user understands, and is better if it can be described in a natural language that is easy to understand. "Non-trivial" refers to the process of data mining is usually non-linear, difficult to be discovered, and can be summed up after continuous comparative analysis using dedicated massive data mining tools. A complete database knowledge discovery process first needs to analyze the requirements, clearly and clearly locate the business problems, and determine the purpose of knowledge discovery. Before starting a project, the final goal of knowledge discovery must be clarified through communication with target users and relevant industry experts [5].

2.2 Data Preprocessing

The main purpose of data preprocessing is to improve the quality of the data set and reduce the complexity of the data. In order to obtain similar meteorological conditions, the weather types are fuzzy identified and classified. Considering the limited conditions of the detection facilities of domestic PV power plants and the universality of the methods, the relevant weather type information predicted by the China Meteorological Forecasting Website is used as the identification factor of the weather type. Since these five factors have different dimensions and the magnitude difference is also large, the five factors can be fuzzyly assigned. In this way, there is a weather type tag every day. In the photovoltaic power generation database, the weather type information can be fuzzy classified and marked above, and stored together with the power generation amount in the photovoltaic power generation database, which facilitates the subsequent data screening and processing [6].

After the previous data preprocessing and classification data sources, the sample information with similar weather types is obtained, but the number of samples obtained in this way is still relatively large and not accurate enough. The gray correlation theory can well select the sample weather forecast websites with high similarity and high

correlation with the prediction ports from these samples to obtain the weather information of the predicted mouth. If the minimum temperature of the predicted port is 60 ° C, the maximum temperature is 25 °C, the relative humidity is 61%, the wind speed is 4, and the weather comprehensive information is cloudy, and the value is set to 3. Let the weather data comparison sequence be t, and obtain the following matrix formula;

$$(t_1, t_2, t_3 \ldots\ldots, t_n) = \begin{bmatrix} t_1(1) + t_3(3) \ + \ \mathrm{t}_n \\ t_2(2) \ + \ t_4(4) \ + \ \mathrm{t}_n \end{bmatrix} \tag{1}$$

In formula (1), t represents the meteorological data sample, and the correlation coefficient is calculated by the matrix formula, and the calculation result is as follows:

$$h(y) = \frac{\min\limits_{i}}{|h(k) - h_i|} \tag{2}$$

Equation (2) is the correlation coefficient result, where h represents the resolution coefficient, y represents the sequence vector, and k is the meteorological information. When the actual current is too large, the photovoltaic cell characteristic curve shifts to the second quadrant, and when the avalanche breakdown voltage is reached, the current is too fast, so this situation should be avoided. That is, when the photovoltaic cell flows excessively, it operates in the second quadrant, which is a drawback of the single-diode circuit, and the photovoltaic cell becomes a load-type power source. The current of the photovoltaic cell suddenly becomes very large in the case of avalanche breakdown. In practical applications, in order to avoid this, a parallel bypass diode is often required. That is, when a diode of a certain group of photovoltaic cells is connected in parallel, when the voltage across a photovoltaic cell becomes a negative value, the parallel diode is turned on, thereby preventing the photovoltaic module from becoming a load-bearing power source and absorbing the external energy, and avoiding the formation of hot spots [7].

3 Realizing Regional Photovoltaic Power Generation Data Mining Algorithm

3.1 Output of Association Rules

In terms of integrity and consistency analysis, traditional data mining algorithms provide support for scheduling methods such as peak and frequency adjustments, and then based on data preprocessing and classification data sources to clarify the range of association rules output. Regional photovoltaic power generation is mainly composed of *CM* junction, upper and lower electrode plates, surface passivation layer and reflective layer. It relies mainly on the photovoltaic effect to convert solar energy into electrical energy. The principle of power generation is as follows:

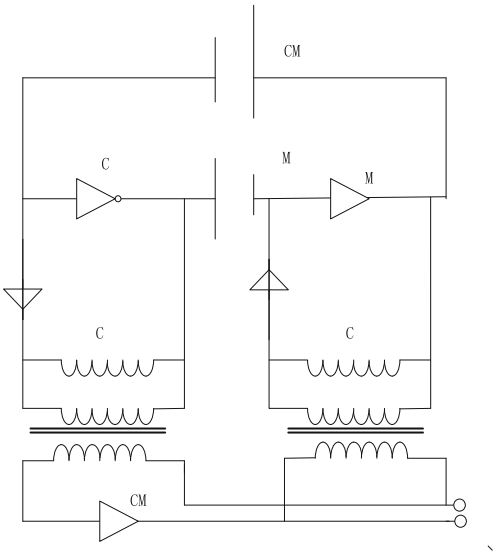

Fig. 1. Power generation principle

Figure 1 shows the principle of power generation. When sunlight is applied to the upper surface of the panel, the *CM* junction will absorb solar radiation energy to generate electron-hole pairs. Due to the influence of the internal electric field, electron-hole pairs will be separated, in which positively charged holes flow from the *M* region to the *C* region, and negatively charged electrons flow from the *C* region to the *M* region. By the flow of electron-holes, positive and negative charges are respectively concentrated near the boundary of the two electrodes, thereby generating a photo-generated electromotive force which is opposite to the electromotive force of the *CM* junction. When the *CM* junction is open, the junction current and the photo-generated current have the same size. Therefore, the two sides of the *CM* junction will produce a stable potential difference between the *C* region and the *M* region, which is the open circuit voltage of the photovoltaic cell. When we connect the *CM* junction to the circuit, under continuous illumination of the mouth, the *CM* junction is equivalent to the power supply [8], and the positive and negative charges continuously flow to form a current, which can convert the solar radiant energy into electrical energy. Expand the system space by limiting the output range of data association rules [9].

3.2 Implicit Data Extraction

A certain data set is satisfied for a given data set based on association rules. First, the data set is constructed into a cluster tree, and then decomposed according to the hierarchy of the cluster tree. The bottom-up and top-down constitute two decomposition sequences of hierarchical decomposition, and complete the function of data mining. The clustering formula is as follows;

$$b_e(l,s) = \left\{ w \cdot \exp\left(\beta \frac{-a(c_l, c_s)}{\varphi} \right) \right\} \tag{3}$$

In Eq. (3), b represents the distance between c_l and c_s, l represents the row of the matrix, s represents the column, and a represents the set of edges. The advantage of this formula is that it can be constructed by hierarchical clustering to form different levels of structure. The disadvantage is that the time complexity is large, and the choice of the decomposition point has a certain influence on the clustering effect. Furthermore, time series analysis is used to deal with the problem of poor clustering effect [10]. Time series analysis determines the degree of similarity between samples by calculating the distance between each sample. The closer the distance between the two samples, the higher the degree of similarity. The time series analysis formula is as follows:

$$b_e = f\left(\frac{-k}{\varphi} (b_l, b_s) \right) \tag{4}$$

In formula (4), b represents a data sample, f represents cluster value, k represents a cluster position, and b_l and b_s represent rows and columns of clusters. The sum of the squares of the total distances of all classes is as small as possible. The larger the value of k, the more classes the sample divides, and the f-value will monotonically decrease as the value of k increases [11]. When the value of k is small, an increase in the number of categories will cause the value of f to decrease rapidly. However, as the value of k increases to a certain value, the change in f value will gradually become flat until the value of k and the total number of samples. Time series analysis is fast and easy to implement. It is efficient when dealing with massive data, and has a certain degree of flexibility. The time complexity of the algorithm is close to linear.

The reason why the PV module array can output electric energy is to convert the energy radiated by the sunlight on the surface into DC electric energy through the photovoltaic effect, and then output the DC power to the battery or the inverter. Photovoltaic array is the most important component in photovoltaic power generation system. In the process of use, the power generation array should be connected in series and parallel according to the required voltage level and power level to meet the power generation needs. The time series analysis is used to decompose the data set and complete the data mining algorithm. In the process of data mining, the attributes of some data are related to time and will change with the passage of time. The time interval of these data can be constant or variable. Time series analysis is mainly to achieve the mining of patterns, similarity search and analysis of data trends [12]. The data mining algorithm flow is as follows (Fig. 2):

In the first step, after the requirements are clarified, the knowledge base data needs to be prepared. Through the integrity and consistency analysis, the data noise is removed and the missing attributes of the data are filled, and the invalid data is also deleted. In the second step, the preprocessor task is to record the data in the standard logger from the initial and final stages of port scanning of a source IP. When a certain log file is determined, the record type, destination IP address, and port are recorded and analyzed. The concept of port scanning is to complete the task of attempting a TCP

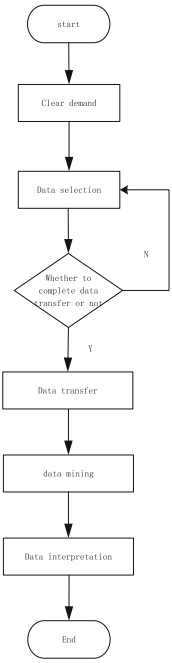

Fig. 2. Classification data source process

connection for more than P ports under the time limit of T (seconds), and it can be said that UDP packets are transmitted to more than P ports under the limitation of time T (seconds). The third step is to transform or unify the data into a form that is easy to mine. After completing the data preparation work, the next step is data mining. The fourth step is to analyze and process the data set according to the different needs of the customer and select the appropriate data mining method for the specific situation. In the fifth step, according to certain evaluation criteria, meaningful models and knowledge should be selected from the results of the mining, and visualization and knowledge expression techniques should be used to express knowledge that can be understood by users in natural language [13].

4 Experimental Results

In order to verify the effectiveness of the regional photovoltaic power generation data mining algorithm, the algorithm was tested. The simulated attack software used in the experiment, which can generate as many kinds of attack data, is an effective tool for testing. Network packets that are used to accurately and accurately record the actual attack process for each type of attack in the experiment. After encapsulation and automatic processing, an attack file is formed, thereby constructing a huge attack library to detect various attack recognition strategies. Simulate attack packets are sent with simulated attack software and saved to a log file. A total of about 2,000 attack

packets were recorded to measure the effectiveness of system pre-detection. The system's pre-detection engine discards packets that are considered normal to reduce unnecessary abuse detection. The test environment is as follows:

Table 2. Test environment

Type of attack	Number
smurf	21561
neptune	26564
back	4123
Satan	4562
ipsweep	2651
portsweep	2654
warezclient	5654
teardrop	5265
Pod	322
land	235

Table 2 is a test environment for testing regional photovoltaic power generation data mining algorithms and traditional data mining algorithms. The test results are as follows:

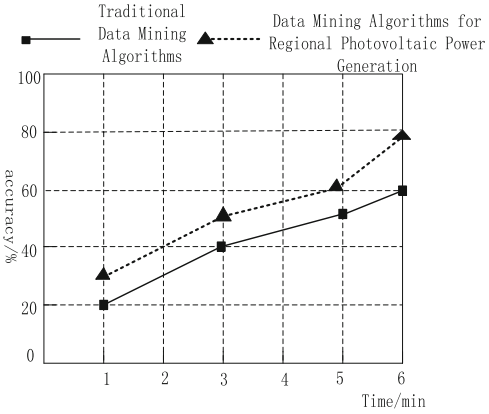

Fig. 3. Comparison of experimental results

Figure 3 is a comparison of experimental results, comparing traditional data mining algorithms with regional photovoltaic power generation data mining algorithms. It also compares the adaptive runtimes. As can be seen from Fig. 3, when the amount of data is small, the regional PV data mining algorithm does not show obvious advantages, and its running time is even slower than the traditional frequent item set mining algorithm. However, with the increase of data volume, the regional photovoltaic power generation data mining algorithm gradually shows its advantages, especially when the data volume

is too large, the traditional frequent item set mining algorithm can not complete the calculation. However, regional photovoltaic power generation data mining algorithms make full use of their parallel advantages, and their speed is not affected by excessive data factors. In addition, after solving the problem of load imbalance, it can be seen that the running time of the regional photovoltaic power generation data mining algorithm is better than the common algorithm. It proves that the proposed regional photovoltaic power generation data mining algorithm has strong stability.

5 Conclusion

Traditional data mining algorithms have the following shortcomings: poor applicability, high false negative rate or false positive rate, too single detection, lack of extensive testing and analysis of rules, and excessive reliance on their own rule base. These deficiencies have not been able to cope with today's increasingly complex network environment. The regional photovoltaic power generation data mining algorithm can store all the relationships between the data items, and then split the entire database into smaller pieces of data for mining processing. The most powerful advantage of this algorithm is that it effectively reduces the number of scans and improves the efficiency, so this is often used for dynamic feature extraction data.

References

1. Wang, H., Li, F., Zhang, L., et al.: Research and application of big data mining technology in photovoltaic power prediction. Hebei Electr. Power Technol. **12**(2), 123–143 (2018)
2. Ge, J.: Research on large data mining algorithm for periodic performance of warships under load environment. Ship Sci. Technol. **23**(22), 121–123 (2017)
3. Xiaobo, Y.: Research on Data Mining Algorithms Based on Projection Pattern Support Set [J]. Computer Applications and Software **34**(7), 273–276 (2017)
4. Lv, Y., Huang, L.: Research and implementation of data mining algorithms based on Hadoop framework in Hongbo. Large data environment. Electron. Design Eng. **25**(7), 241–244 (2017)
5. Zheng, Z., Wu, W., Li, H.: Data mining algorithms based on CLUSTER optimization. J. Harbin Commercial Univ. (Nat. Sci. Ed.) **33**(6), 329–330 (2017)
6. Shibing, B.: Research on data mining algorithms based on neural network and particle swarm optimization. Laser J. **38**(3), 188–192 (2017)
7. Cheng, Z., Li, S., Han, L., et al.: Research on the prediction method of photovoltaic array power generation based on data mining. J. Solar Energy **38**(3), 726–733 (2017)
8. Yang, J., Lin, X., Yao, Q.: Study on wind power generation system based on hydraulic constant rense speed control. J. Xi'an Polytech. Univ. **2018**(05), 574–580 (2018)
9. Yongwen, Y., Fanyue, Q., Zaiqin, T., et al.: Data mining and analysis method for operation log of photovoltaic power plant. Sci. Technol. Econ. Guide **34**(2), 114–145 (2017)
10. Luo, R., Wang, Z., Wu, T., et al.: A pollution diagnosis method for photovoltaic modules based on large data mining of solar power regulation. CN107065829A [P] **33**(2), 134–135 (2017)

11. Ding, M., Li, C., Li, H., et al.: Fault detection method for photovoltaic inverters based on massive data mining. Electr. Autom. **33**(3), 154–165 (2018)
12. Gong, S., Pan, T., Wu, D., et al.: Research on the method of filling the missing photovoltaic data in microgrid based on MCMCMC. Renew. Energy **33**(3), 453–465 (2018)
13. Wang, H., Li, F., Zhang, L., et al.: Research and application of big data mining technology in photovoltaic power prediction. Hebei Electr. Power Technol. **12**(2), 345–367 (2018)