DOI: 10.23977/acss.2023.071106
ISSN 2371-8838 Vol. 7 Num. 11

Research on the influencing factors of wind power generation based on clustering and decision tree

Yaoqi Tan

School of Electrical Engineering, Chongqing University, Chongqing, 400044, China

Keywords: Decision Tree, K-means, Wind Power Generation

Abstract: In this paper, the decision tree method is utilized to explore the influencing factors of wind power generation. This paper innovatively utilizes a combination of clustering and decision tree algorithms for data analysis. Firstly, the samples are categorized into three categories, high wind power generation, medium wind power generation and low wind power generation using K-means algorithm. Then, a decision tree model was applied to each category to obtain the proportion of feature importance. The results show that the key factors affecting wind power generation include motor torque, blade angle, electrical resistance and generator temperature. Compared to the traditional Adaboost algorithm, the new algorithm has a mean square error of no more than 3% and a coefficient of determination (R^2) greater than 0.78. Compared to the Adaboost algorithm, the new algorithm has a 2.671% lower mean square error and an improved R2 of 0.135, which suggests that the new algorithm is more reliable in predicting wind power generation. Future research directions, this study can be extended by considering more factors that affect wind power generation, such as wind speed, wind direction, and air density.

1. Introduction

In recent years, with the increasing prominence of environmental issues and the growing demand for renewable energy, wind power as a green and environmentally friendly energy source has attracted much attention. As an important device for converting wind energy into electricity, wind turbines play an important role in the optimization and improvement of energy structure. In fact, wind turbines have become an important trend in the future development of energy and electricity.

However, wind turbines are often affected by environmental conditions during operation. Due to the combined effect of various factors, the power generation of wind turbines is unstable, especially in the low wind speed section, the power generation is consistently low, which has become the focus of attention and difficulty in the current wind power business. In order to solve this problem, the relevant staff urgently need to consider the factors affecting the power generation of wind turbines and make a series of technical improvements from these factors to increase the power generation of wind turbines[1-2].

The literature[3] based on grey correlation theory proves that turbulence intensity and rotational inertia are the dominant factors affecting the performance of maximum power point tracking control of wind turbines. In order to explore the influence factors of wind turbine power, this thesis will be based on the decision tree method. Decision tree is a commonly used machine learning method that is able to construct a tree structure by analyzing and learning the features in a dataset in order to predict or classify new data[4]. By applying the decision tree method, we can deeply explore the factors affecting the power of wind power generation, including but not limited to wind speed, wind direction, air temperature, humidity, atmospheric pressure and other factors.

This thesis aims to conduct in-depth research and investigation on the power influencing factors of wind turbine by establishing a decision tree model. By analyzing and mining a large amount of measured data, we hope to find the degree of influence of key factors on the power of wind turbine and provide feasible technical improvement solutions for the wind power industry, so as to improve the power generation and stability of wind turbine. This will be of great significance in promoting the development of renewable energy as well as optimizing and improving the energy structure.

2. Exploration of the influence factors of wind power generation power based on Decision Tree

2.1 Main ideas

In this paper, the data size is excessive and the data features are not apparent. Thus, initially, the clustering method is employed for dimensionality reduction. By utilizing the K-means algorithm, the data samples are categorized into three groups. Subsequently, decision tree regression is conducted for each category to investigate the impact of individual factors on wind power generation size.

2.2 K-means Model

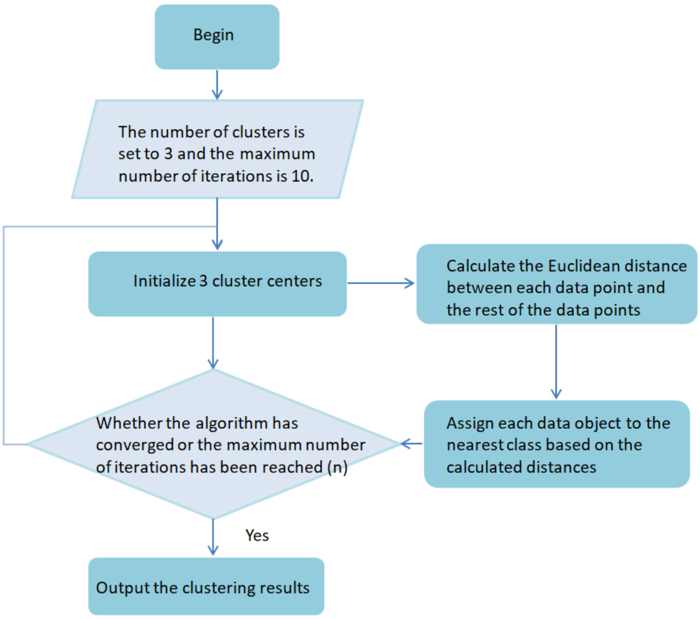
Begin
The number of clusters is set to 3 and the maximum number of iterations is 10.
Initialize 3 cluster centers
Calculate the Euclidean distance between each data point and the rest of the data points
Assign each data object to the nearest class based on the calculated distances
Whether the algorithm has converged or the maximum number of iterations has been reached (n)
Yes
Output the clustering results

Figure 1: K-means algorithm flowchart

The K-means algorithm is a distance-based clustering algorithm that uses distance as a measure of similarity. The closer the distance between two objects, the more similar they are considered to be. The algorithm aims to create compact and independent clusters by selecting k initial class clustering centers. These initial centers have a significant impact on the clustering results[5].

In the first step of the algorithm, k objects are randomly selected as the initial clustering centers, representing individual clusters. Then, in each iteration, the algorithm assigns each remaining object to the nearest cluster based on its distance from the cluster center. This process continues until all data objects have been examined.

After each iteration, new cluster centers are computed. The algorithm converges when the value of J, which represents the evaluation index, does not change before or after an iteration. The specific process is illustrated in Figure 1.

2.3 Decision Tree Model

The decision tree is a tree-like structure that begins with a root node. It tests the data samples and divides them into different subsets based on the results. Each subset forms a child node. This method classifies data by following a set of rules, providing insights into the values obtained under specific conditions[6]. The specific process is illustrated in Figure 2and 3.

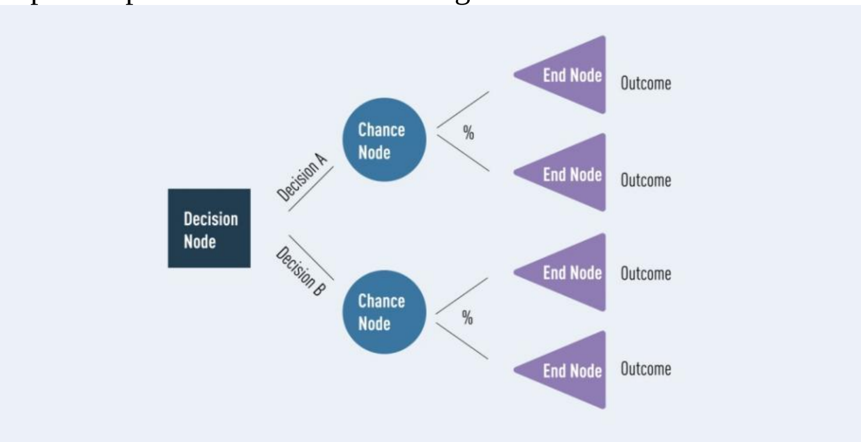

Figure 2: Decision Tree Diagram

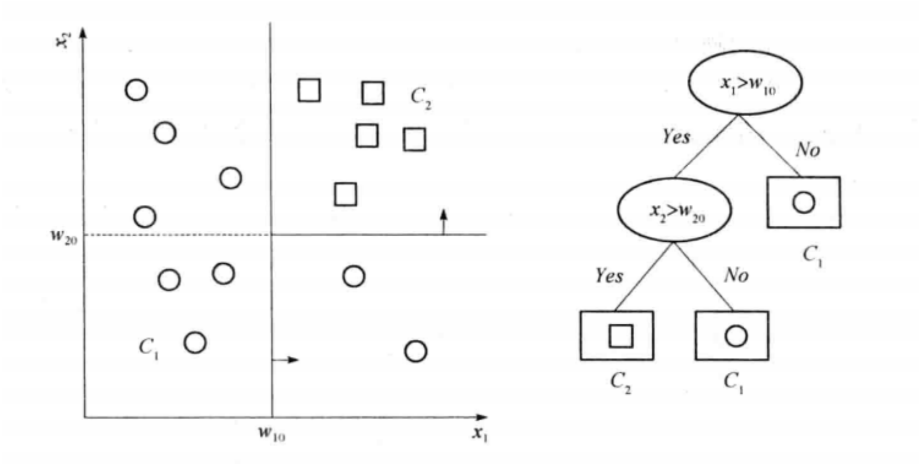

Figure 3: Decision Tree Schematic

The ID3 algorithm chooses attributes to test using a criterion known as gain, which relies on the notion of entropy in information theory. Let S be a set of s data samples. Assuming the category label attribute has m distinct values, define m distinct categories Ci (i=1, ..., m). Let si represent the number of samples in category Ci. The expected information needed to classify a given sample is calculated using the following equation:

$$l(s_1, s_2,\cdot, s_m) = -\sum_{i=1}^{m} p_i \log_2(p_i) \tag{1}$$

where pi=si/s is the probability that any sample belongs to i. Note that the logarithmic function has a base of 2. The reason for this is that the information is encoded in binary. Let attribute A have v different values {a1, a2,..., ax}. Attribute A can be used to partition S into v-enterprise subsets {S1, S2,...,Sy} ,where the samples in Sj have the same value ai(j=1,2,..., v) on attribute A. Let sij be the number of samples of class Ci in subset Sji. The full or information expectati to v-enterprise subsets {S1, S2,...,Sy} ,where the samples in Sj have the same value ai(j=1,2,..., v) on of the subset divided into subsets by A is given by the following equation:

$$\mathrm{E(A)} = \sum_{\mathrm{j=1}} \left(\frac{(\mathrm{s_{1j}+s_{2j}+\cdots+s_{mj}})}{\mathrm{s}}\right)^{*} \mathrm{l}\left(\mathrm{s_{1j}} + \mathrm{s_{2j}} + \cdots + \mathrm{s_{mj}}\right) \tag{2}$$

The smaller the entropy value, the higher the purity of subset division. For a given subset Sj, its information expectation is given by the following equation:

$$\mathrm{I}\left(\mathrm{s_{1j}} + \mathrm{s_{2j}}+\ldots+\mathrm{s_{mj}}\right) = -\sum_{\mathrm{i=1}}^{\mathrm{m}} \mathrm{p_{ij}}\log_2\left(\mathrm{p_{\ddot{y}}}\right) \tag{3}$$

where pij=sij/sj is the probability that a sample in Sj belongs to Ci.

The information gain that will be obtained by branching on attribute A is given by the following equation:

$$\mathrm{Gain(A)} = \mathrm{I}(\mathrm{s_1}, \mathrm{s_2}, .., \mathrm{s_m}) - \mathrm{E(A)} \tag{4}$$

C4.5 The algorithm uses the gain ratio as a criterion for selecting attributes at all levels of the decision tree.

$$\mathrm{SplitInformation}(A,S) = -\sum_{i=1}^{c} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|} \tag{5}$$

$$\mathrm{GainRatio}(A,S) = \frac{\mathrm{Gain}(S,A)}{\mathrm{SplitInformation}(S,A)} \tag{6}$$

3. Example analysis

3.1 Data composition and sources

To assess the algorithm's effectiveness, the dataset includes variables such as wind speed, atmospheric temperature, generator temperature, rotor torque, windmill generated power, and more. Dataset is from a wind farm in Fuzhou, China.

3.2 Quantitative metrics for model evaluation

$$\mathrm{MSE} = \frac{\sum_{\mathrm{i=1}}^{\mathrm{n}}(\mathrm{y_i} - \hat{\mathrm{y}}_{\mathrm{\imath}})^2}{\mathrm{n}} \tag{7}$$

$$\mathrm{R}^2 = 1 - \frac{\sum_{\mathrm{i}} (\hat{\mathrm{y}}_{\mathrm{\imath}} - \mathrm{y_i})^2}{\sum_{\mathrm{i}} (\bar{\mathrm{y}}_{\mathrm{\imath}} - \mathrm{y_i})^2} \tag{8}$$

3.3 Experimental results and analysis

First, the samples were divided into three classes using the K-means algorithm, and the clustering results are displayed in Figure 4.

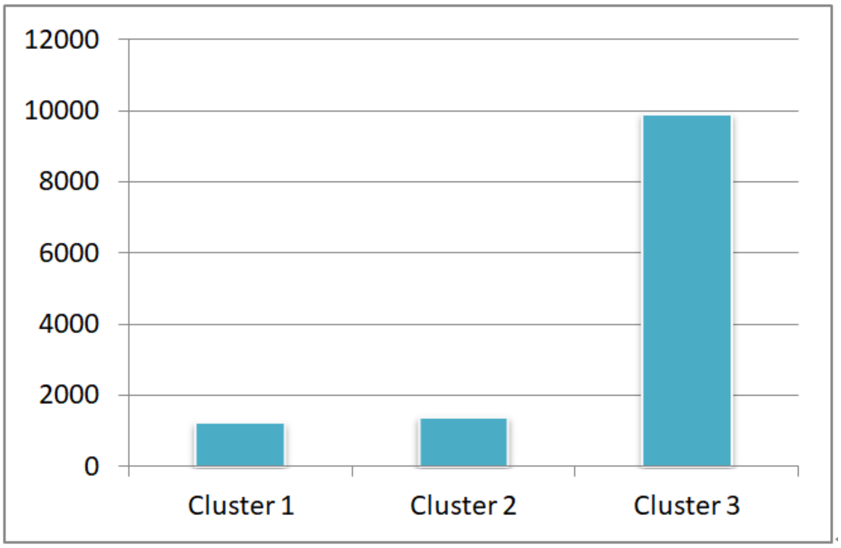

Figure 4: Clustering results graph

The K-means algorithm classifies the samples into high wind power, medium wind power, and low wind power.

Then, the decision tree model is used to learn each category and determine the proportions of feature importance. The proportions of feature importance for each category are then averaged to create the final plot of independent variable importance.

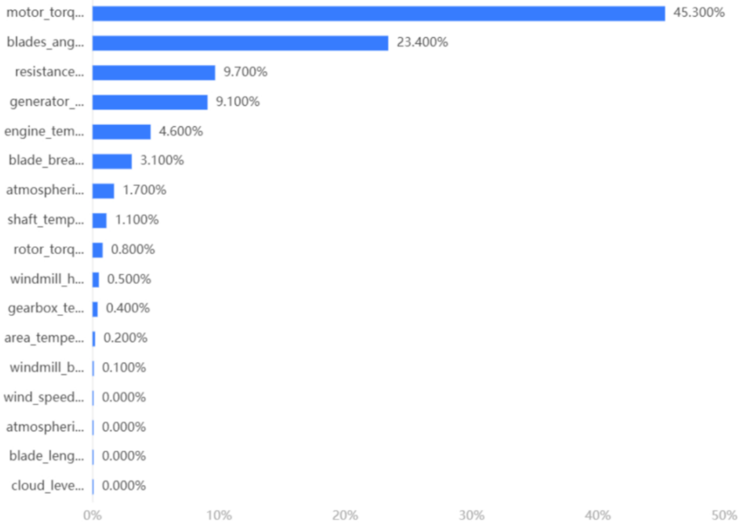

Figure 5: Characteristic Importance Chart.

From the figure 5, it is evident that motor_torque plays a significant role in determining the size of wind power, with an importance ratio of 45.300%. This is followed by blades_angle, which has an importance ratio of 23.400%. Additionally, resistanceo and generator_temperature have importance ratios of 9.7% and 9.1% respectively. The remaining indicators have minimal influence on wind power.

For comparison, the unclassified samples are trained using the Adaboost algorithm to determine the corresponding feature importance. To evaluate the performance of the new method, Mean Square Error (MSE) and Coefficient of Determination (R^2) are used for comparative analysis. The results of this analysis are presented in Table 1.

Table 1: Comparison among forecasting performances

Cluster11DT		Cluster21DT		Cluster31DT		Adaboost	
MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2
1.107	0.86	2.926	0.78	1.688	0.821	3.778	0.725

From Table I, it is evident that both algorithms are viable analysis methods, as the mean square error is below 4%. When compared to the Adaboost algorithm, the decision tree algorithm, based on clustering results, shows a decrease in MSE values from 3.778% to 1.688%, 2.926%, and 1.107%. Additionally, the R2 values have increased to 0.821, 0.78, and 0.86. This new algorithm enhances prediction accuracy and offers greater reliability.

4. Conclusions

In this paper, we begin by creating a K-means clustering model to group the samples. The K-means algorithm classifies the samples into high wind power, medium wind power, and low wind power. Subsequently, we construct a decision tree model to assess the importance of the indicators. Analysis results from examples demonstrate that the new algorithm's regression mean square error is below 3%, with an R2 value exceeding 0.78. When compared to the traditional Adaboost algorithm, the regression mean square error is reduced by a maximum of 2.671%, while the R2 is enhanced by a maximum of 0.135. The results indicate that this method has high accuracy in predicting wind power generation and is effective in practice. Therefore, this research demonstrates high feasibility and practicality. In the future, further exploration of the application scope of this method can be conducted, such as applying it to other renewable energy fields like solar power and hydro power. Additionally, combining this method with other machine learning algorithms can be attempted to improve prediction accuracy. Furthermore, the use of deep learning models for wind power generation prediction can be considered to further enhance prediction precision. Overall, this research provides a new approach for the renewable energy field with broad prospects for application.

References

[1] Han Xiaoxiao, Zhang Xiaohua, Bu Bing, et al. Factors affecting wind power generation and optimization strategy [J]. Power Equipment Management, 2023(10):63-65.

[2] Peng L , Jing-Chao W , Bin L ,et al.Research on Data Mining of Wind Disaster of Power Transmission Line Based on Clustering Analysis[C]//2019 6th International Conference on Information Science and Control Engineering (ICISCE).2019.DOI:10.1109/ICISCE48695.2019.00098.

[3] Gu Yundong, Ma Dongfen, Cheng Hongchao. Electricity load forecasting based on similar data selection and improved gradient boosting decision tree [J]. Journal of Power System and Automation, 2019, 31(5):64-69.

[4] Yang Juanxia. Research on Vibration Fault Diagnosis of Wind Turbine Generators Based on Clustering and Decision Tree Algorithms [J]. Power Equipment Management, 2023 (16): 100-102.

[5] Liu, Hongfu, Chen, Junxiang, Dy, Jennifer, Fu, Yun.Transforming Complex Problems Into K-Means Solutions[J]. Ieee Transactions On Pattern Analysis And Machine Intelligence, 2023, 45(7):9149-9168.

[6] Ma Lichuan, Peng Jiayi, Pei Qingqi, Zhu Haojin. An efficient decision tree privacy classification service protocol [J]. Journal of Communication, 2021, 42(8):80-89.