Anomaly Monitoring of Power Characteristic of Wind Turbine based on Multi-dimensional Clustering Method

Jianlou Lou*, 1, Hui Cao1, Hongjian Zheng2 and Jizhe Xiao1

1School of Information and Engineering, Northeast Electric Power University, Jilin 132012, China
2PetroChina jilin Petrochemical Company Power First Plant, Jilin 132022, China
1loujianlou@qq.com, 2jh_zhenghj@petrochina.com.cn

Abstract: Aiming at the limitation of traditional abnormal monitoring method of wind turbine. In this paper, we propose a multidimensional clustering method called WPMCLU. Firstly, traverse the FP-Tree storage structure to find subspace instead of APRIORI self-connection mode. Secondly, define K Gauss models and identify clusters in each subspace. At last, according to the parameter Eq specified by user to divide normal and abnormal cluster of each subspace, remove redundancy and recognize abnormal data. And the method we propose runs on the Spark platform in order to easily extend to large data set of wind power. In experiment, the method we propose in this paper has a high recognition rate compares with CLIQUE, K-Means and DBSCAN algorithms.

Keywords: multidimensional clustering; large data set; abnormal monitoring; Spark Platform; wind turbine

1 Introduction

Wind energy is a non-polluting renewable energy and the wind power is one of the fastest developing energy sources in the electricity field [1]. However, wind turbines in addition to facing the problems of serious climate issues, such as: wind, rain and snow. Also faces the problems of its own control system, such as: yaw system, variable pitch control, etc. [2]. This will lead to a deviation between the actual power generation and the expected power generation. And unable to evaluate the performance of wind turbine well. So how to fully and accurately monitor the abnormal running state can effectively save the cost and achieve the maximum power output.

The traditional anomaly detection methods are mainly from two aspects to monitor the running state of the wind turbines: Identification of abnormal operation status of the wind turbine by machine learning and control chart based on two dimensional data of wind speed and power [3-7]. Based on the method of machine learning to construct the power curve model accurately. Making the abnormal point deviate from the normal range of the power curve to achieve the abnormal recognition [8-10]. However, all abnormal points can’t be identified only through two dimensional attributes, especially adhesion to normal data. And it is not easy to extend to large data set.
The multidimensional clustering algorithms are presented as follows: CLIQUE, ENCLUS and MAFIA algorithms [11-13] divide the multidimensional data space into a number of grid cells according to the threshold and select dense grid cells based on APRIORI self-connection method to find the maximal subspaces. CBF, CLTree and DOC [14-16] algorithms also divide the data space as above, and the dense cells are stored in a special structure to map the maximum subspace. However, the accuracy of clustering results is not high and can form density connectivity among the clusters.

2 Anomaly Recognition based on WPMCLU Method

2.1 Identifying the Maximum Subspace

Firstly, the data need to discretize. The data range of each dimensional attribute is divided into multiple equal intervals, the intervals in the same dimensional are labeled with the same item and with the digital identification among the different intervals. The data points in the same unit are represented by the interval identification of current unit.

In frequency calculation part. We firstly calculate the frequency degree of each unit. The data set is divided into high frequent units and low frequent units. And delete the low frequent units. Secondly, remove digital identification from the remaining high frequent degree units. The number of units with the same attribute identification is calculated to account for the percentage of all units. And the frequency of each units are sorted from high to low. The formula is as follows:

\[
\text{frequency degree}(I) = \frac{\text{number}(I)}{\text{number}(D)}
\]  

(1)

In formula: number(I) is the number of discrete data for the calculation of the frequency in the data set; number(D) is the number of discrete data for the entire data set; frequency degree(I) is the percentage of discrete data for calculation in the whole discrete data set.

The processed data is stored in the form of FP-Tree storage structure [17] and traverse FP-Tree to find the maximum subspaces. The special steps are as follows:

Step1: Traverse the leftmost leaf node of FP-Tree and according to the formula (1) to calculate the current node frequency. If the frequency is greater than the threshold, the node as the initial node and traverse the parent until root. The attributes corresponding to the nodes on the same tree trunk are as the maximum space. If the frequency is less than the threshold value, traverse the next leaf node. And so on, access to all leaf nodes.

Step2: Delete the leaf node that the frequency is less than the threshold value. And the parent node as a new leaf node. According to the above traverse process (That is: step1) to search the subspaces.

Step3: Identify the subspace is mainly to find the initial node. The initial node can’t be the traversed node, and can't be the nearest node to the root node. To delete the traversal process if meet the above situation.

Step4: This process is performed recursively until all subspaces are sought.
2.2 Identification Clusters in Subspace

In this part, we define the K Gauss models [18] to identify the clusters in each subspace. Then K mixed Gauss models formula is as follows:

\[ P_x(x) = \sum_{k=1}^{K} \pi_k N(x; \mu_k, \Sigma_k) \]

(2)

In formula: K is the number of Gauss model; \( \pi_k \) is the select weigh; \( \mu_k \) is the variance; \( \Sigma_k \) is the mean value.

In each subspace, the data set is partitioned into K clusters. Firstly, initialize the initial parameters of each Gauss model. Secondly, the weights of each Gauss model are calculated. Perform this two steps iteratively up to the parameter convergence.

2.3 Clustering Merging and Anomaly Recognition

According to the parameter Eq to divide the clusters in each subspace. Preliminary the large and density clusters are defined as normal data, others are defined as abnormal data. Because in the subspace clustering part, we can set up a relatively large K value, that is, there are more clusters in each subspace. At this time, we need to set up a small parameter Eq. to reduce the likelihood that the normal data is wrongly identified as abnormal data. Although this will result in only a part of abnormal data points can be identified in each subspace. But through the superposition of multiple subspaces anomaly recognition, can well monitor the abnormal operation state of wind turbine.

The normal clusters and abnormal clusters of all subspaces are added separately, and remove redundancy. And then divided into normal and abnormal clusters. If there are still redundant data in normal and abnormal clusters, then call the Gauss model (K=2) again to classify the redundant data.

2.4 Abnormal Detection of Wind Turbine based on WPMCLUS

The specific identification processes of anomaly data of the wind turbine are as follows:

(1) Partition the multidimensional data of wind power. And according to the 3.1 step to calculate table F;
(2) The table F data is stored in the form of FP-Tree storage structure and based on 3.1 steps to traverse the FP-Tree. The corresponding properties of the nodes accessed by the same traversal process constitute the largest subspace;
(3) To define K mixed Gauss models, and to cluster the data sets in each subspace;
(4) According to the parameter Eq, the number of points in a cluster larger than parameter Eq are defined as normal data. And others are defined as abnormal data.
(5) The normal and abnormal data of all the subspaces are added separately, and remove redundancy. If the normal and abnormal data still exist redundancy, then call the Gauss model again to identify the redundant data (K=2). At last, output the result.
3 Instance Verification

3.1 Abnormal Detection based on WPMCLU

In this part, the method we propose runs on the Spark platform framework [19] and we identify the normal data and abnormal data based on the WPMCLU method. Based on the two properties of wind speed and power, rotor speed and wind speed, blade angle and power to do two-dimensional scatter plot. As shown in figure 1:

![Scatter Plot](image1.jpg)

Fig.1. Clustering based on WPMCLU method

From Figure 1: It can be seen that the multidimensional clustering method proposed in this paper can effectively identify the abnormal data point adhesion with normal data.

3.2 Comparison of Algorithms

In order to show that the proposed method has advantages compared with other methods, we compare the method we propose with CLIQUE and K-Means algorithms. As shown in Figure 2:

![Scatter Plot](image2.jpg)

Fig.2. Algorithm Comparison

In figure 2(left), the abnormal detection based on CLIQUE algorithm. We can see that the multidimensional clustering method based on grid can’t accurately identify most of the abnormal points. In figure 2(right), the abnormal detection based on K-Means algorithm. We can see that the abnormal detection of SCADA data of wind turbine is not suitable for the multidimensional clustering method in the full dimensional space.
In order to further illustrate the effectiveness of method we propose in this paper. Identify the abnormal data of full year based on the method we propose, DBSCAN, CLIQUE and K-Means algorithms. And a total of 52081 records. In this paper, the artificial identification method defined as: Based on the artificial experience, the output power in the normal range under the same wind speed are normal data, others are abnormal data. The abnormal recognition rate as shown in Table 1:

Table 1. 4 Algorithms to Identify the Abnormal Data Volume Table

<table>
<thead>
<tr>
<th></th>
<th>Artificial</th>
<th>WPMCLU</th>
<th>DBSCAN</th>
<th>CLIQUE</th>
<th>K-Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abnormal data</td>
<td>29965</td>
<td>28957</td>
<td>5678</td>
<td>6724</td>
<td>13956</td>
</tr>
<tr>
<td>Recognition rate</td>
<td>\</td>
<td>96.7%</td>
<td>19.0%</td>
<td>22.4%</td>
<td>46.6%</td>
</tr>
</tbody>
</table>

As can be seen from the table 1, based on the method we propose to identify the amount of abnormal data is close to the amount of abnormal data identified by artificial statistics and has a very good recognition rate.

4 Conclusion

Aiming at the limitation of the traditional method of monitoring the power characteristic of wind turbine, we propose a multidimensional clustering method WPMCLU. And compare with CLIQUE, K-Means and DBSCAN algorithm has a high recognition rate. At the same time, the method runs on the Spark platform to show a good scalability to large data sets in order to meet the future development needs of wind power field data processing.

In addition, the method we propose can evaluate the performance of the wind turbine, reduce the maintenance cost and maintenance difficulty, to achieve the maximum output of wind turbine. However, the WPMCLU clustering method is based on the data points in each subspace to cluster, so the time complexity is higher in subspace clustering part. How to reduce the time complexity of this part will be the task of the next research.

References