Exploratory analysis of a real wind turbine dataset, using AI tools to cluster and classify data, for condition monitoring and fault detection

Análise exploratória de um conjunto de dados de turbina eólica real, usando ferramentas de IA para agrupar e classificar dados, para monitoramento de condições e detecção de falhas

Received: 2023-05-20 | Accepted: 2023-06-23 | Published: 2023-06-30

ABSTRACT

In recent years there has been an increase in wind generation, driven by environmental factors and the incentive offered for the development of clean and sustainable technologies for energy generation. However, due to the rapid growth of this technology, concerns about the safety and reliability of wind turbines are increasing, especially due to the associated risks and financial costs. Therefore, health monitoring and fault detection for wind turbines has become an important research focus. Thus, the...
The objective of this work was to realize an exploratory study of real data from a wind turbine, using AI tools that help to group the different behaviors, according to the similarity of resources and characteristics of the data. For this, unsupervised learning methods were used to cluster the data and a model was proposed to train and test, using a multilayer perceptron network, to classify these clusters. The differential of this work is the use of real data from CHESF’s wind turbines. Another important contribution is in relation to permanent magnet wind turbines, as there are not many studies in this field, therefore a great potential to be explored.

**Keywords:** Wind turbine; Maintenance; Fault detection; Cluster; Multilayer perceptron.

---

**RESUMO**

Nos últimos anos houve um aumento da geração eólica, impulsionado por fatores ambientais e pelo incentivo oferecido para o desenvolvimento de tecnologias limpas e sustentáveis para geração de energia. No entanto, devido ao rápido crescimento desta tecnologia, as preocupações com a segurança e confiabilidade das turbinas eólicas são cada vez maiores, especialmente devido aos riscos associados e aos custos financeiros. Portanto, o monitoramento da integridade e detecção de falhas para turbinas eólicas tornou-se um importante foco de pesquisa. Assim, o objetivo deste trabalho foi realizar um estudo exploratório de dados reais de uma turbina eólica, utilizando ferramentas de IA que ajudem a agrupar os diferentes comportamentos, de acordo com a similaridade de recursos e características dos dados. Para isso, métodos de aprendizado não supervisionado foram usados para agrupar os dados e um modelo foi proposto para treinar e testar, usando uma rede perceptron multcamadas, para classificar esses agrupamentos. O diferencial deste trabalho é a utilização de dados reais dos aerogeradores da CHESF. Outra contribuição importante é em relação aos aerogeradores de imã permanente, pois não há muitos estudos nessa área, portanto um grande potencial a ser explorado.

**Palavras-chave:** Turbina eólica; Manutenção; Detecção de falha; Agrupamento; Perceptron multcamadas.
INTRODUCTION

The energy consumption global has increased exponentially and significant efforts have been made by most developing countries to attenuate the impacts of climate change by maximizing clear energy use that to minimize greenhouse gas emissions (BASHIR; BASHIR, 2022).

The Intergovernmental Panel on Climate Change (IPCC) commented on the future risks emanating from climate change, stating:

Continued emission of greenhouse gases will cause further warming and long-lasting changes in all components of the climate system, increasing the likelihood of severe, pervasive, and irreversible impacts for people and ecosystems. Limiting climate change would require substantial and sustained reductions in greenhouse gas emissions which, together with adaptation, can limit climate change risks (PACHAURI; MEYER, 2014).

Wind energy is a virtually carbon-free and pollution-free electricity source, with global wind resources greatly exceeding electricity demand. Accordingly, the installed capacity of wind turbines (WTs) grew at an annualized rate of bigger then 20% from 2000 to 2019 and is projected to increase by a further 50% by the end of 2023 (PRYOR; BARTHELME R.J.AND BUKOVSKY; OTHERS., 2020).

Due to these external variances, WTs undergo constantly changing dynamics and local loads, resulting in a large variation in operating conditions that lead to intense mechanical stress (WANG et al., 2022). Many of the already installed WTs are aging, driving the growing maintenance and repair market, along with the demand for the development of new maintenance and repair technologies. As the demand for wind power continues to grow at exponential rates, maintenance will be a permanent factor related to costs, and it can directly influence energy prices and the competitiveness of renewable energy (MISHNAEVSKY; THOMSEN, 2020). Maintenance (O&M) activities are a critical aspect of reliability. As a matter of fact, the global O&M market is projected to grow to USD 27.4 billion by the year 2025. Maintenance is one of the leading costs in the total expenditure of a wind farm project and, if not effective, can cause drastic losses in energy production due to downtime (ARTIGAO et al., 2018).

Considering this context, it is important to detect failures before they occur to avoid greater financial losses. Therefore, knowing the standard behavior of a WT and being able to classify health conditions, so that the O&M can act at the right time, justify studies like this one.

In this work, an initial exploratory study was realized using real WT data, from a CHESF wind farm, selecting a period that, due to a previous mapping of the O&M reports, contained failures. The objective was to analyze data and identify behavior patterns that can be classified as: normal operation, failure and pre-failure. This application is only possible due to the existence...
of failure labeling in the equipment data, but on the other hand, the pre-failure condition does not present labeling, justifying the use of unsupervised learning methods to highlight these possible patterns.

Using AI tools, a model was built to analyze a dataset of a real WT, separate these data into clusters, using non-supervised models and classify them into different behavior patterns using multilayer perceptron (MLP), for later application.

The results obtained in the training and testing phase were above 98%, which was already an excellent result. It is important to expand the application of the model using other datasets and deepen the studies of the behavior of the WTs, to confirm that the patterns found, really can be classified as the possible states of health of the WTs.

RELATED WORK AND CONTRIBUTIONS

As already mentioned, environmental concerns, global warming and other factors have awakened governments, large companies and the academic community to the benefits of wind generation. Therefore, they are investing a lot of money to develop new technologies in this area and the number of studies, documents and patents related to wind turbines has only grown.

In August 2022, a patent search was conducted to evaluate the technology of interest, systems, methods and devices for wind turbines. Figure 1 details the annual distribution of patent applications regarding the described technology; the first applications were found in 1982. The number of patent document applications showed growth over the analyzed period, indicating a continued strong interest in the development of new technologies for wind energy (BARBOSA et al., 2023).

Figure 1 – Time analysis of patent documents. First year of priority.

Source: Elaborated by the authors (2023)
Reder, Yürüşen e Melero. (2018) used a fault dataset of modern wind turbines and applied supervised and unsupervised data mining techniques to obtain logical interconnections between faults and environmental data.

Turnbull et al. (2019) built a dataset of bearing failures based on wind turbine real data resulting from crossing between failures identified in O&M reports and SCADA data, then they trained machine learning algorithms to classify bearings as healthy or not.


Hsu et al. (2020) used statistical control and machine learning techniques to diagnose failures in wind turbines and determine maintenance needs, based on the analysis of real data collected from 2015 to 2017 in Taiwan and data provided by Taipower professionals, about possible causes of failures.

El-Menshawy, Gul e El-Thalji (2021) analyzed a two-year dataset of a SCADA wind turbine, using machine learning algorithms such as: boosted decision tree, decision forest and K-means cluster, in Azure ML Learning Studio.


This work is interesting because it uses real data from a permanent magnet wind turbine from the Casa Nova A wind farm. There are not many studies on permanent magnet wind turbines. In addition, the results of this work will be used for various applications, such as modeling condition monitoring systems and helping the maintenance of CHESF’s wind farms.

**MATERIALS AND METHODS**

This study was based on real operational data from Wind Generator 18, with permanent magnet, located at state of Bahia, Brazil, in the Casa Nova A windfarm (Latitude: 09º 09’ 43” S / Longitude: 40º 58’ 15” W), and made use of data processing, exploration, and interpretation approaches, as well as the application of machine learning techniques, for grouping, selection of attributes and classification.

The proposed computational model was constructed from seven essential steps: acquisition of data from the wind turbine using SCADA, the pre-processing step, which will be described in the next topic, the clustering of the database, then a selection of attributes, followed by training, and classification, whose results will be applied for fault detection. Figure 2 demonstrates this flow of steps for modeling.
Figure 2 – Computer modeling procedure.

Source: Elaborated by the authors (2023)

Dataset

The dataset used was generated from wind turbine 18 of the Casa Nova A windfarm, obtained from the SCADA – Supervisory Control and Data Acquisition, provided by CHESF. This database was the result of a fragment of almost four months, measurements every ten minutes of a period chosen precisely because it contains some failure. This information was previously mapped in the maintenance reports provided by Goldwind, the company responsible for the maintenance of this wind turbine. This data needed to be pre-processed, Figure 3 resumes this data pre-processing. The files obtained from SCADA are in multiple CSV files that need to be unified, for this a translation of the data that are in Chinese is carried out, in addition to synchronization and use of metrics such as: average, mode and median, to add and remove duplicate values.

Figure 3 – Data pre-processing.

Source: Elaborated by the authors (2023)
The result is a single CSV file, but to be able to apply the AI techniques, it needs to be further refined. Starting from 423 attributes, a deeper analysis of the existing data was necessary to work only with the most relevant ones. In the end, 106 attributes remained: 91 of average values, 4 booleans, 1 indication of time, another of failure and 9 percent attributes.

**Clustering**

Clustering, in the data mining aspect, is considered an active method of grouping data into many collections or clusters according to similarity of data features and characteristics (EZUGWU et al., 2022).

Unsupervised learning corresponds to statistical methods that extract meaning from data without training a model on labeled data. This model also builds a data model but does not distinguish between response and predictor variables. This model can be used to create a predictive rule in the absence of a labeled response, so this model can be seen as an extension of exploratory data analysis. Unsupervised techniques will allow examining and analyzing variables and discovering existing relationships (GÉRON, 2019).

Expectation-maximization EM algorithm is an iterative method which alternates between two steps, expectation (E) and maximization (M). For clustering, EM makes use of the finite Gaussian mixtures model and estimates a set of parameters iteratively until a desired convergence value is achieved. This method only infers about dataset, pointing out the groupings, according to the similarities founds (DEMPSTER; LAIRD; RUBIN, 1977).

K-means is a popular algorithm was developed and perfected by researchers Lloyd (1957) and MacQueen (1967), its concept is easy to understand and can be used in grouping themes, identification of patterns or anomalies (LLOYD, 1957; MACQUEEN, 1967). The goal of k-means is to obtain a partition that minimizes the squared error between the mean of a cluster and the observations within that cluster. Different from EM in the k-means is necessary to previously define the desired number of clusters. Next steps:

- Random initialization of k-centroids based on the number of previously selected clusters.
- Calculation of the distance between the observed data and definition of the cluster based on the smallest relative distances.
- Calculation of new centroids for each cluster generated in the previous step.
- Repetition the last two steps until convergence, that is, there no change in the positioning of the centroids.

The biggest computation cost is in the step where are calculated the distances between the k-centroids and all N points observed, resulting in an amount of processing N x k. For the calculation of the distances, the k-means method usually uses these functions: Euclidean Distance or Manhattan Distance.
Selecting Attributes

In this study, two ways were used to select the main attributes, using select attributes tools and the exploratory analysis, performing visual comparisons and previous maintenance knowledge.

The Weka machine learning software has attribute selection tools, which choose the main attributes by evaluating the value of a subset of attributes considering the individual predictive capacity of each resource along with the degree of redundancy between them (FRANK et al., 2016).

An exploratory analysis of the attributes was realized, looking for correlations between them, in addition to maintenance knowledges. For example, the Figure 4 shows very similar attributes, then is possible choose only one.

Figure 4 – Similar Attributes.
Classify

It was decided to use a multilayer perceptron network to build the pattern classification and recognition model.

The classification process of grouped data using Multi-Layer Perceptron (MLP) is subdivided into two steps: training (learning phase) and data classification (test). The complete dataset, with the labels, is divided into two parts, one for the training step and the other for the classification step. In the training step, the classification model is built from the labeled data. Then, the classification step is executed, which analyzes the dataset extracting the labels, to try to associate each element with the class label that it belongs to. Since each element's class label is provided to the classifier, this step is known as supervised learning.

RESULTS AND DISCUSSIONS

Table 1 presents the results of clustering the complete dataset, with 106 attributes and 16,414 instances, using EM and k-means, choosing: Euclidean Distance and Manhattan Distance. The processing time of EM was much longer than k-means times, but in EM, it is not necessary, to previously define the number of clusters.

<table>
<thead>
<tr>
<th></th>
<th>EM</th>
<th>K-means (EuclideanDistance)</th>
<th>K-means (ManhattanDistance)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clustered Instances</td>
<td>4.521</td>
<td>1.128</td>
<td>1.496</td>
</tr>
<tr>
<td>Percentage</td>
<td>28%</td>
<td>7%</td>
<td>9%</td>
</tr>
<tr>
<td>Clustered Instances</td>
<td>1.809</td>
<td>13.300</td>
<td>12.943</td>
</tr>
<tr>
<td>Percentage</td>
<td>11%</td>
<td>81%</td>
<td>79%</td>
</tr>
<tr>
<td>Clustered Instances</td>
<td>10.084</td>
<td>1.986</td>
<td>1.975</td>
</tr>
<tr>
<td>Percentage</td>
<td>61%</td>
<td>12%</td>
<td>12%</td>
</tr>
</tbody>
</table>

Source: Elaborated by the authors (2023)

Starting from the exploratory analysis and with the objective of removing similar attributes, it was possible to verify an interesting result with the reduction of 106 attributes to 72, the Table 2 shows the results of clustering with 72 attributes and 16,414 instances, using EM and k-means, choosing: Euclidean Distance and Manhattan Distance.
### Table 2 – Clusters results and time take to build models.

<table>
<thead>
<tr>
<th></th>
<th>EM</th>
<th>K-means (EuclideanDistance)</th>
<th>K-means (ManhattanDistance)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Clustered Instances</td>
<td>Clustered Instances</td>
<td>Clustered Instances</td>
</tr>
<tr>
<td>0</td>
<td>1.928 12%</td>
<td>1.116 7%</td>
<td>1.136 7%</td>
</tr>
<tr>
<td>1</td>
<td>11.406 69%</td>
<td>13.299 81%</td>
<td>13.277 81%</td>
</tr>
<tr>
<td>2</td>
<td>3.080 19%</td>
<td>1.999 12%</td>
<td>2.001 12%</td>
</tr>
</tbody>
</table>

Time: 102.85 seconds 0.16 seconds 0.30 seconds

Source: Elaborated by the authors (2023)

Removing some similar attributes, much better results were achieved in clustering, shorter times and much closer results comparing EM with k-means.

Figure 5 shows the selected attributes from the complete dataset, with 106 attributes and from the dataset with 72 attributes, after exploratory analysis. Both selections were performed using the system’s selection tools, the difference was the input dataset, but the result is the same.

**Figure 5 – Attribute selection.**

Using MLP for classification, it started with the complete dataset with 108 input attributes (108 for being the post-grouping dataset), then the dataset with 74 was used, but it was not successful. So, based on the selections of the main attributes, the attributes were removed until this case:
Table 3 – Attributes selected to MLP classify.

<table>
<thead>
<tr>
<th>Attributes Selected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instance_number</td>
</tr>
<tr>
<td>average-speed-of-pitch-1-c</td>
</tr>
<tr>
<td>average-yaw-speed-c</td>
</tr>
<tr>
<td>instantaneous-value-of-average-speed-of-generator-rpm</td>
</tr>
<tr>
<td>min-inclination-angle-c</td>
</tr>
<tr>
<td>average-temperature-of-the-side-compensation-capacitor-c</td>
</tr>
<tr>
<td>enable-power-curve-generation-flag</td>
</tr>
<tr>
<td>the-failure-statistics-enable-the-flag</td>
</tr>
<tr>
<td>%cp1</td>
</tr>
<tr>
<td>Cluster</td>
</tr>
</tbody>
</table>

Source: Elaborated by the authors (2023)

The selection used to perform the classification using MLP was with 10 input attributes, shown in Table 3. Hall et al. (2009) defines the number of hidden layers, using this formula:

\[ N_{\text{hidden}} = \sqrt{N_{\text{in}} \times N_{\text{out}}} \]

The MLP architecture obtained was this: 10 input attributes and 3 outputs, therefore the indicated number of hidden layers is root 30, which is approximately five. Therefore, two inner layers of five neurons each were chosen. The Table 4 contains the list of the network hyperparameters used as input to the simulator and Figure 6 shows the MLP architecture.

Figure 6 – MLP architecture.

Source: Elaborated by the authors (2023)
Table 4 – Network Hyperparameters.

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of layers</td>
<td>4</td>
</tr>
<tr>
<td>Number of hidden layers</td>
<td>2</td>
</tr>
<tr>
<td>Number of hidden units (for each hidden layer)</td>
<td>5/5</td>
</tr>
<tr>
<td>Batch size</td>
<td>100</td>
</tr>
<tr>
<td>Epochs</td>
<td>500</td>
</tr>
<tr>
<td>Activation function</td>
<td>Sigmoid</td>
</tr>
<tr>
<td>Learning rate</td>
<td>0.3</td>
</tr>
<tr>
<td>Momentum</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Source: Elaborated by the authors (2023)

For ME clustering, with 16,414 instances selected to compound the training dataset, the model could classify 16,091 instances correctly, achieving an accuracy of approximately 98.0322%. The confusion matrix is shown in Figure 7. This figure also shows the confusion matrix to k-means clustering, using EuclideanDistance and the same 10 attributes. This model accuracy was approximately 99.9147% and presents the confusion matrix to k-means clustering, using ManhattanDistance, the same 10 attributes and the model accuracy was approximately 99.9086%.

Figure 7 – Confusion Matrix.

![Confusion Matrix - ME Clustering](image1)

![Confusion Matrix - K-means Clustering (EuclideanDistance)](image2)

![Confusion Matrix - K-means Clustering (ManhattanDistance)](image3)

Source: Elaborated by the authors (2023)
CONCLUSION

The current environmental condition requires urgent changes, so the development of new technologies in clean sources of energy generation is essential. In addition, remote monitoring of wind turbines is very important, as this equipment is usually installed in isolated and difficult to access places.

In this context, this study focused on the analysis of data from a wind turbine and on the treatment of these data applying AI tools.

Interesting results were found analyzing SCADA data with clustering tools. It was verified that k-means processing time is better than EM, but it is necessary to define the desired number of clusters.

From the result obtained by the EM, 3 clusters were found, which indicate possible behaviors of: normal operation, pre-failure and failure, however further analysis will be necessary, using more data and a previous mapping to guarantee more conclusive results.

The exploratory analysis of the attributes and the selection using tools, demonstrate that there is a lot of redundancy in the input dataset, so it is important to invest some time to analyze the data and selecting main attributes to save on processing time.

The model built for training and recognition of the clusters obtained an excellent result, with an accuracy above 98%. It is important to extend the analyzes using new input data to be sure of the effectiveness of the model.

As already mentioned, this work is introductory and will contribute to the development of a condition monitoring system, fault detection and recommendations for the maintenance team of CHESF’s wind farms.
REFERENCES


