# Fault Prediction and Health Management Based on Big Data and Its Application in Guide Vanes of Hydropower Units

Dong Chen

Gongzui Hydro Power Plant, Guoneng Dadu River Basin Hydropower Development Co., Ltd., Leshan, China

12011065@ceic.com

**Abstract.** The development and application of a new generation of artificial intelligence technology equipment has accumulated a large amount of data, driving the fault prediction and health management (PHM) into the industrial big data era. Combined with the functional role, structural composition and working characteristics of the equipment, it is urgent to analyze the equipment big data and realize the condition monitoring, abnormality warning, fault diagnosis, life prediction and intelligent maintenance of the equipment. The abnormal data detection model of GCDEG is proposed. While analyzing the technical connotation, development status and application of the proposed algorithm, the characteristics of big data in equipment industry, analysis methods and the difficulties and doubts in its work are discussed. Taking the guide vane, a key equipment of hydropower unit, as an example, the anomaly detection technology is discussed from the perspective of industrial big data. This could help to provide certain reference for researchers in related fields.

**Keywords:** Big data; GCDEG; Anomaly detection; Fault detection, Health management.

## 1 Introduction

With the growing demand for efficient and stable power from customers, future hydropower equipment maintenance systems will become more complex, posing significant challenges to the way hydropower equipment is managed today[1-2]. The maintenance system will provide a large number of monitoring services for more hydropower equipment, which will be used to support the growing service types and service quantities, and its service architecture should have more flexibility[3-4]. To this end, new architectures and technologies such as big data-based equipment failure prediction and health management are introduced into the maintenance system to reduce the reliance on manual labor and achieve rapid deployment of new services to meet the unattended demand of hydropower plants[5-6].


B. Siuta-Tokarska et al. (eds.), *Proceedings of the 2024 2nd International Conference on Management Innovation and Economy Development (MIED 2024)*, Advances in Economics, Business and Management Research 300,
https://doi.org/10.2991/978-94-6463-542-3_24

Hydro generator set is a device that converts the potential energy of water into the kinetic energy of hydraulic turbine, and then changes the kinetic energy into electric energy through the rotation of generator. Hydro generator set has many parts, huge volume and high integration degree, it is a whole composed of many fixed parts and rotating parts that cooperate and coordinate with each other[7]. The sound indexes issued by the hydraulic turbine, generator and many rotating parts under a certain fixed operating condition are relatively fixed. If the sound and vibration signals under different operating conditions of the unit are collected, stored, and analyzed and processed using identification and diagnostic equipment, it is possible to track and judge the operating status of the hydro generator set equipment. This paper proposes a fault prediction and health management method based on big data to get an abnormality detection model. When an abnormality occurs in a component can be found at the first time and take corresponding processing measures to prevent defects from becoming too large and guarantee the safe power generation of the equipment.

However, these technologies increase the complexity of the monitoring system structure at the same time also make the equipment management and maintenance more difficult, the probability of failure of the detection system itself will increase, affecting the provision of alarm services, and cannot meet the current hydropower station managers of the dependence on the maintenance system and high demand. Traditional manual and semi-automatic fault management, processing faults require manual intervention, resulting in fault processing results susceptible to human influence day. The emergence of big data-based equipment failure prediction and health management simplifies the manual involvement part[8-9]. Among them, the purpose of model training is to establish a normal operation model through the historical operation data of the equipment; self-optimization is to automatically optimize the system parameters after the initial white configuration of the model is completed; and anomaly detection is responsible for the automatic detection of faults or anomalous signals.

## 2 Theory and Algorithms

In order to further improve the effectiveness of the analysis of anomalous behaviors with the existing GMM clustering algorithms, for the existing GMM clustering algorithms need to know the number of clustering clusters as well as the disadvantage of the number of iterations, the GMM algorithm of joint Graph-Community-Detection and EM is proposed, which is referred to as the GCDEG algorithm (Graph-Community-Detection-EM- GMM). The algorithm is mainly divided into two steps: ① using Graph-Community-Detection algorithm for the first clustering of the data obtained from key devices such as the guide leaf, to complete the selection of the k centroids and the subsequent initialization of the parameters of the Gaussian model; ② by the first step of obtaining the k clusters of cluster centers, the EM algorithm iteratively using the parameters of the GMM, so as to analyze the data in a grounded manner. The guide leaf big data analysis method is shown in Fig. 1.

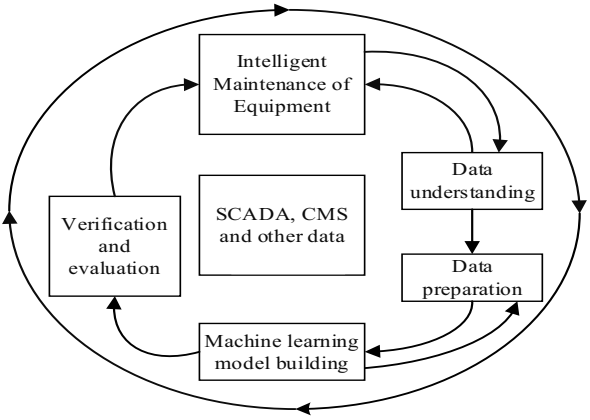


**Fig. 1.** Big data analysis method of the guide leaf

## 2.1 GCDEG Algorithm

The GCDEG algorithm is constructed based on the Graph-Community-Detection clustering algorithm and the EM algorithm as the basic GMM clustering algorithm. The whole set of algorithms first seeks the appropriate number of clusters k and the initial parameters of the subsequent GMM algorithms by performing the first clustering with the Graph-Community-Detection clustering algorithm, and then the EM algorithm is used to iteratively solve the parameters of the GMM model. The basic flow of algorithm construction is refined as follows:

Let the guide leaf information data $x$ have n features, and the definition $\boldsymbol{x} = [x_1, x_2, \cdots, x_n]^T$. Assume that $x$ obeys the distribution of Equation (1):

$$\boldsymbol{x} \sim \sum_{i=1}^{k} \delta_i \cdot p(\boldsymbol{x} \mid \boldsymbol{\mu}_i, \boldsymbol{T}_i) \tag{1}$$

The distribution consists of k Gaussian components, where $\delta_i$ denotes the probability from the $i$th Gaussian distribution, satisfying the relationship of Equation (2):

$$\sum_{i=1}^{k} \delta_i = 1, \quad \delta_i \geq 0 \tag{2}$$

$p(\boldsymbol{x} \mid \boldsymbol{\mu}_i, \boldsymbol{T}_i)$ is the probability density distribution function of x, which is defined in Equation (3):

$$p(\boldsymbol{x} \mid \boldsymbol{\mu}_i, \boldsymbol{T}_i) = \frac{\exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)^T \boldsymbol{T}_i^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i)\right]}{(2\pi)^{\frac{n}{2}} \mid \boldsymbol{T}_i \mid^{\frac{1}{2}}} \tag{3}$$

where $\boldsymbol{\mu}_i$ is the mean vector, $\boldsymbol{T}_i$ is the covariance matrix, and $(\cdot)^T$ is the transpose operation. Equation (3) states that the parameters of this distribution consist of the mean vector and the covariance matrix. For analytical purposes, the hidden variable $p(y_j = i)$ is introduced. Notice that $p(y_j = i) = \delta_i$, which means the probability that sample $\boldsymbol{x}_i$ is from the *j*th Gaussian distribution component. When the condition of Equation (4) is satisfied, $\boldsymbol{x}_j$ is regarded as being from the *i*th Gaussian component:

$$\max p(y_j = i), \quad j = 1, 2, \cdots, k \tag{4}$$

When the hidden variables are known, the Gaussian mixture clustering model divides the data set $D = \{x_1, x_2, \cdots, x_m\}$ into *k* classes of data. At this time, in order to obtain the model parameters $(\delta_i, \boldsymbol{\mu}_i, \boldsymbol{T}_i)$, the maximum likelihood estimation method is needed to solve the parameters. The likelihood function of dataset *D* is shown in Equation (5):

$$L(D) = \prod_{j=1}^{m} [\sum_{i=1}^{k} \delta_i \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{T}_i)] \tag{5}$$

The above derivation yields the computation process of $\boldsymbol{\mu}_i$, $\boldsymbol{T}_i$ and $\delta_i$. When these 3 parameters are computed, a posteriori probability $p(y_j = i \mid \boldsymbol{x}_j)$ is updated in turn. According to the Bayesian formula, Equation (2) can be obtained:

$$p(y_j = i \mid \boldsymbol{x}_j) = \frac{p(y_j = i) \cdot p(\boldsymbol{x}_j \mid y_j = 1)}{\sum_{i=1}^{k} \delta_i \cdot p(\boldsymbol{x} \mid \boldsymbol{\mu}_i, \boldsymbol{T}_i)} \tag{6}$$

The process of the EM algorithm is divided into 2 steps: after initializing the model parameters, the probability of each sub-model is first calculated according to Equation (6). Secondly, the 3 model parameters are updated according to Equation (1) and iterated until the likelihood function grows slowly or no longer grows. If the parameters of initialization are not properly selected, it will lead to a large number of iterations, which directly affects the efficiency of the clustering process. Usually the initialization method adopts random selection. That is, k points are randomly selected as the clustering center, and the distance between the remaining data and these k points is calculated, and the data are divided into categories according to the criterion of the closest distance. After completing the division, all the data in the k categories are obtained and the initial parameters $(\delta_i, \boldsymbol{\mu}_i, \boldsymbol{T}_i)$ are calculated. After that, the iteration of GCDEG algorithm is performed.

### 2.2 Applications

The data cleaning work allows the constructed model to more easily capture the interrelationships between the features of the guide vane health operation stages, and by learning and recognizing such relationships; the model can more accurately monitor the health status of the guide vane unit. Subsequently, the cleaned historical data of the healthy operation of the guide vane unit is normalized and used to train the neural network, which uses the characteristics of the network to maximally retain the information in the original data while performing dimensionality reduction on the original data features. The individual features were visualized in Fig. 2.

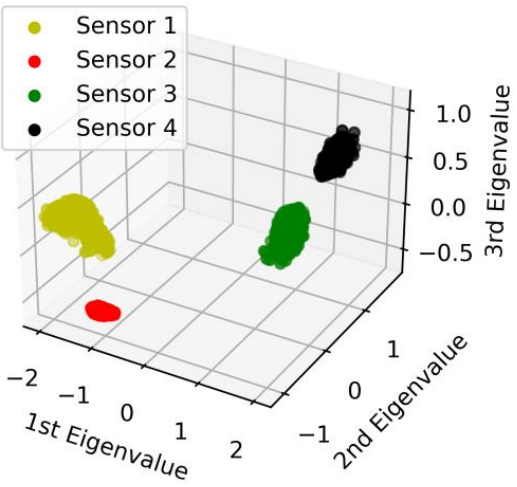


**Fig. 2.** Vector distribution of the combined feature

In order to verify the effectiveness of the proposed method, due to the lack of data labels on the data of hydropower units, this paper takes the post-cleaning data of the unit as the basis and manually simulates the abnormal data for cleaning verification. In order to ensure that the data set covers all the working condition areas, this paper randomly samples the data of each working condition area as the basic data set, and according to the mean and variance of each working condition area, under the principle of data obeying the same probability distribution within the same working condition area, the anomalous data are manually simulated. Three kinds of mixed data sets are constructed respectively, and the proportion of their abnormal data is shown in Table 1. As shown in Table 1, GCDEG multi-scale data cleaning doubles the recognition rate of anomalies compared with single GCD or EG data cleaning, which indicates that the method proposed in this paper can significantly improve the cleaning quality. Meanwhile, the misidentification rate of GCDEG is lower than 1/3 of GMM, which indicates that the method is more capable of retaining healthy data and reducing the omission of important information.

**Table 1.** Results of manual simulation of data set

| Method | Data set | Anomalies/pc | Misidentified points/pc | Recognition rate/% | Misidentification rate/% |
|---|---|---|---|---|---|
| GCDEG | 1 | 157 | 25 | 32.9 | 15.9 |
| | 2 | 323 | 80 | 40.3 | 24.7 |
| | 3 | 590 | 130 | 39.3 | 22.0 |
| GMM | 1 | 369 | 10 | 69.8 | 2.7 |
| | 2 | 724 | 23 | 72.9 | 3.1 |
| | 3 | 996 | 39 | 73.6 | 3.9 |

## 3 Conclusion

With the rise of renewable energy penetration, the frequency of changes in the operation mode of the power system increases significantly, and the operation modes of two adjacent days may show completely different patterns, which indicates that more flexible resources and dispatchable means are needed in the operation and scheduling of the power system in order to cope with the power balance problems brought about by frequent changes in the operation mode.

## References

1. Zhong, X., Wang, Y., Ming, G., Liu, J. (2020) Data-driven fault diagnosis method based on compressed sensing and improved multiscale network. IEEE Trans. Ind. Electron., 67: 3216–25. DOI: 10.1109/tie.2019.2912763.
2. Liu, R., Yang, B., Zio, E., Chen, X., (2018) Artificial intelligence for fault diagnosis of rotating machinery: a review Mech. Syst. Signal Process., 108: 33–47. DOI: 10. 1016/ j.ymssp.2018.02.016.
3. Qian, Q., Qin, Y., Luo, J., Wang, Y., Fei, W., (2023) Deep discriminative transfer learning network for cross-machine fault diagnosis. Mech. Syst. Signal Process., 186: 109884. DOI: 10.1016/J.YMSSP.2022.109884.
4. Jiang, W., Luo, J., (2022) Graph neural network for traffic forecasting: a survey. Expert Syst. Appl., 207: 117921. DOI: 10.1016/J.ESWA.2022.117921.
5. Li, T., Zhao, Z., Sun, C., Yan, R., Chen, X., (2021) Domain adversarial graph convolutional network for fault diagnosis under variable working conditions. IEEE Trans. Instrum. Meas., 70: 1–10. DOI: 10.1109/TIM.2021.3075016.
6. Nadkarni, S., Prügl, R. (2020) Digital transformation: a review, synthesis and opportunities for future research. Management Review Quarterly, 71:233-241. DOI: 10.1007/s11301-020-00185-7.
7. Verhoef, P. C., Broekhuizen, T., Bart, Y., et al. (2021) Digital transformation: a multidisciplinary reflection and research agenda. Journal of Business Research, 122: 889-901. DOI: 10.1016/j.jbusres.2019.09.022.
8. Chen, Z., Jiamin, X., Peng, T., Yang, C., (2022) Graph convolutional network-based method for fault diagnosis using a hybrid of measurement and prior knowledge. IEEE Trans. Cybern., 52: 9157–69. DOI: 10.1109/TCYB.2021.3059002.
9. Zhu, Y., Zhuang, F., et al. (2021) Deep subdomain adaptation network for image classification. IEEE Trans. Neural Netw. Learn. Syst., 32: 1713–22. DOI: 10.1109/TNNLS. 2020.2988928.