Mechanical equipment fault detection applying data mining technology

Yiqun Liu¹ᵃ, Xiaogang Wang¹ᵇ*, Xiaoyuan Gong¹ᶜ, Hua Mu²ᵈ

¹School of Intelligent Manufacturing and Control Engineering, Shanghai Polytechnic University, Shanghai, 201029, P. R. of China
²Wuhan KM Information Technology Co., Ltd., Hubei, 430076, P. R. of China
ᵃe-mail: lyq1995@live.com,ᵇ*Corresponding author: xgwang@sspu.edu.cn
ᶜe-mail: 841952969@qq.com,ᵈe-mail: yqyqgo@163.com

Abstract—Mechanical equipment fault detection in manufacturing enterprises has always been an important link in the manufacturing process. Along with the computer technology, artificial intelligence technology and various sensors are widely used in manufacturing industry, the amount of data produced by manufacturing machinery and equipment at all stages of the production process is also increasing rapidly. It is particularly important to analyze the data generated by these devices for fault detection and even prediction, Data mining technology provides advanced analysis methods for this purpose. This paper first introduces the basic concepts of Data Mining, Data Mining process and the key technology of Data Mining, and then focuses on the application of Data Mining in machinery fault detection, finally through the analysis of Data Mining in the field of top class meeting KDD2019, explored the could be applied to machinery fault detection of new technology, forecast the Data Mining technology in the fault detection of possible development trend.

1. INTRODUCTION
With the introduction of intelligent manufacturing, "Industry 4.0", "Made in China 2025" and other concepts, the manufacturing industry in all countries in the world develops rapidly [1], and a large number of intelligent, automatic, networked and digital new manufacturing equipment is used, and massive data information is accumulated [2]. There are a lot of valuable potential information in these continuously generated and accumulated massive data. How to quickly obtain effective data from massive data and provide support for equipment fault detection has become an urgent matter of the moment, while the traditional method of manually processing data is obviously unable to meet the needs. Data mining technology analyzes the data generated by manufacturing equipment, excavates the laws contained therein [3], and analyzes the operation state and operation law of manufacturing equipment based on these laws, so as to achieve the purpose of saving cost, improving efficiency and reducing energy consumption, and gradually make the manufacturing industry intelligent [4].

2. DATA MINING AND ITS MAIN ALGORITHMS
Data mining is simply mining or extracting useful knowledge from a large number of data. On the definition of the concept of data mining has many versions, in general, data mining is the data from a large database to extract implicit, previously unknown, valid, novel and potential application of the knowledge and information, the extracted knowledge representation for the Concepts, Rules,
Regularities, such as form, this definition to define the object of data mining for the database or data warehouse. In addition, some literatures call Data mining as Knowledge Discovery, Knowledge Extraction, Data Archaeology, Data Dredging, Intelligent Data Analysis, and so on [5].

The object of data mining is a very broad, mainly oriented relational database, data warehouse, text (text unstructured or semi-structured data), multimedia data (including images, audio, video, data), Web data and complex type of data (mainly includes spatial database and time sequence database), etc.

The task of data mining is to find patterns from data sets. According to the functional patterns, it is mainly divided into descriptive mining and predictive mining. The emphasis of descriptive mining is to find interesting patterns to describe data and describe general characteristics of data. Predictive mining is predictive inference based on current and historical data, focusing on the behavior of predictive models. At present, the main tasks of data mining are usually divided into data description, model evaluation, prediction, classification, cluster analysis and correlation analysis according to the practical application of the pattern. Data description is an Exploratory Data Analysis (EDA) that enrichs existing Data to give a conceptual description. A high-quality description can usually be implemented in Exploratory Data Analysis, a graphical approach that Exploratory Data can be used to search for patterns and trends [6]. Model evaluation is to establish a model and evaluate the difference between the target variable and the prediction variable through the approximate value of the regular prediction variable of the data set. Simple linear regression or correlation and multiple regression are generally used, and neural network can also be used for model evaluation [7-9]. Classification is similar to model evaluation, except that the target variable of classification is category instead of number. The main methods include KNN, decision tree, neural network, logistic regression, Bayesian network, etc. [10-13]. Prediction is similar to classification and model evaluation, but prediction is mainly aimed at the future situation. General classification and model prediction methods are applicable to prediction analysis. Clustering and classification, clustering, no target variable database objects in the cluster is divided into multiple classes, objects of the same class similarity as large as possible, not the same kind of objects as small as possible, the similarity of more common method used to measure object similarity distance, density and so on, generally use the clustering method with KNN, BIRCH [14], etc. Association analysis is to discover the correlation between data and discover which attributes exist at the same time. Data association is a kind of important discoverable knowledge existing in the database. Generally, support degree and credibility are used to measure the correlation of association rules, and common association algorithms are Apriori algorithm and generalized rule inductive algorithm [15-16].

3. KEY TECHNOLOGY OF DATA MINING

Data mining consists of three stages: data preprocessing, data mining and evaluation and presentation of mining results. Figure 1 shows the entire process of data mining.

3.1. Data preprocessing

Data preprocessing is a very important part in the whole process of data mining. When we integrate the data from various data sources, the data is often incomplete and there are many noises and redundancy. The quality of the data directly affects the reliability of the mining model and the correctness of the decision. For the accuracy of data mining results, data needs to be preprocessed, which mainly includes three forms of data cleaning, data transformation and data dimension reduction.

3.1.1. Data cleaning: The content of data cleaning mainly includes processing missing data, identifying misclassification and identifying outliers. For a data set containing a large amount of data, information loss is very common. Although data with missing information can be processed faster by simply omitting or discarding data, it is likely to result in deviation and data waste. Therefore, some specified constants or estimates based on other features are usually used instead. Finding and correcting misclassification can also improve the accuracy of the results. Outliers are extreme values deviating from other value trends. Some data mining methods will be greatly affected by outliers, which may lead to result errors. Therefore, it is very important to identify outliers.
3.1.2. **Data transformation:** The range of different variables is often very different. For some data mining algorithms, such range difference will lead to the variable with large range will have a bad influence on the result. Therefore, it is necessary to normalize the data in the process of data mining, such as normalizing the data, so as to standardize the impact of each variable on the results.

3.1.3. **Data dimension reduction:** Data sets used in data mining may contain millions of pieces of data and thousands of variables, not all of which are independent and unrelated to each other. In the process of data mining, it is necessary to prevent the correlation between predictive variables from causing instability of the results. Principal component analysis (PCA) [17] is a statistical analysis method that selects a small set of important variables to describe the relevant structure through linear combination of multiple variables. Principal component analysis is used only for predictive variables and not for target variables.

3.2. **Data mining**

Data mining is the most critical link in the whole data mining process, which usually includes four aspects: 1. Selecting and applying appropriate data mining technology; 2. Calibrating model Settings to optimize results; 3. Applying various technologies; 4. In this stage, the algorithm is selected according to the target of data mining, and the corresponding pattern model is mined by analyzing the data.

3.3. **Evaluation and presentation of results**

Model evaluation requires a comprehensive review of the data mining process to determine whether important factors or tasks have been neglected for some reasons. Model representation is visualization, which enables the model to be presented in a friendly manner now available to users. Because the pattern model mined in the second phase is not necessarily meaningful or desirable to the target user. Therefore, the data should be interpreted and evaluated.

This stage is also important to present the target user with the desired, easy-to-understand, visual pattern model.

![Data mining flowchart](image)

Figure 1. Data mining flowchart
4. APPLICATION OF DATA MINING IN FAULT DETECTION OF MECHANICAL EQUIPMENT

Fault detection of mechanical equipment can monitor, diagnose and predict the status and faults of continuously running mechanical equipment, and ensure the safe operation of mechanical equipment, which is of great significance to ensure the safe operation of equipment.

Machine as a dedicated to tunnel large machines, bad work environment, prone to failure, and the hydraulic system is one of the main parts of machine failure, and boring machine hydraulic system components, complex circuit, and sealing, lead to the failure mechanism, influential factors and forms are diverse, the operator is difficult to accurately find out the cause of the problem of fault handling correctly. Because of the complexity of the hydraulic system of road header, the traditional manual diagnostic method is difficult to meet the diagnostic requirements, so the data mining method must be adopted. Literature [18] proposes a fault diagnosis method combining fuzzy theory with BP neural network for the hydraulic system fault of road header, and VC ++ and Matlab are applied to realize the fault diagnosis software. The experiment shows that the system has good effect. Literature [29] describes the state information of the motor equipment by recording 6 attribute data, and extracts its fault diagnosis rules based on rough set idea, so that the system can monitor its production state in real time through the running state of the equipment, detect equipment faults in time, and ensure the normal operation of the production system. Literature [20] for different fault factors have different effect in the process of fault diagnosis, this paper proposes a matrix of weighted association rules algorithm, using fault factors weights to improve the accuracy in fault diagnosis, through the improvement of shear connection reduces the scale of candidate item sets generation, thus improve the efficiency of fault diagnosis. Literature [21] proposed FDC (Fault Detection and Classification) method. First, principal component analysis was used to sort out training data and establish a modular neural network for Fault identification. D-s theory is used for fault classification in the case of uncertainty (such as multiple fault generation or fault transmission). The FDC method based on modular neural network and D-S theory has the ability to quickly identify the abnormal and locate the root component of the abnormal when it occurs, so that relevant personnel can quickly locate the fault component and take corresponding measures to solve it. Literature [22] proposed an improved algorithm of decision tree structure based on variable precision rough set on the basis of the existing decision tree method, which improved the classification accuracy and noise data suppression ability, and applied it to the machinery and equipment of coal plant, and obtained a good effect. In reference [23], a state monitoring system for vibration fault diagnosis of rotating equipment is developed based on the data of large rotating equipment, combined with Apriori algorithm and using B/S structure, which has achieved good results in practical production and application. Literature [24] proposed the visualization of Riemann prevalence and covariance matrix distribution for detection of mechanical equipment anomalies, and applied it to the fault detection of wind turbine gear box, and obtained good results. Literature [25] proposed a black hole particle swarm optimization (BPS-SA) algorithm according to the least square support vector machine and combined with the simulated annealing algorithm, which was greatly improved in the classification speed, accuracy and extreme value problems, and was applied to the fault diagnosis of fan gear box, and obtained good results.

Table 1 shows the application of some data mining techniques in fault detection of mechanical equipment. It can be seen from the table that at present, data mining technology has been applied in many aspects of mechanical equipment fault detection. At the same time, in practical application, a combination of various data mining technologies is generally needed to achieve the expected effect, and targeted improvements should be made to the original data mining technology.

**TABLE I. APPLICATION OF DATA MINING TECHNOLOGY IN FAULT DETECTION OF MECHANICAL EQUIPMENT.**

| Type of mechanical equipment                  | Data Mining technology                      |
|-----------------------------------------------|---------------------------------------------|
| Hydraulic system of road header [18]         | Fuzzy Set and BP Neural Network              |
| Motor equipment [19]                         | Rough set                                   |
| Hydraulic equipment for steel mill [20]      | Matrix weighted association rules            |
5. TRENDS AND CHALLENGES
At the 2019 KDD conference, Shen [26] et al. from Tsinghua University proposed a mobile event prediction framework that integrates spatiotemporal data. This framework has good generalization ability and can be easily transplanted to other fields. In mechanical fault detection, if the fault can be predicted in advance, it will greatly reduce the loss of enterprises.

Google [27] aggregates the features of historical examples to make parking predictions, making full use of various data, the combination of different users and different seasons makes the parking forecast on the Google map more effective. In the future, it may be more effective to detect and predict mechanical equipment faults by combining the personal information of equipment users and other climatic and weather factors.

However, the application of data mining technology in mechanical equipment fault detection still faces many challenges. In terms of technology, the research results involved in this paper are mostly in the experimental stage, and the experimental data are mostly high-quality data screened and reviewed manually. However, many data quality problems are faced in the actual production environment. In addition, a large number of studies are carried out based on offline historical data. The analysis of offline historical data can obtain the operation rules of the equipment, but the data volume must be accumulated to a certain extent, and it is also a difficult problem to use the real-time online data to reflect the latest changing state and situation of the equipment. At the same time, many factors such as the generality, flexibility and robustness of the model and the actual performance of the algorithm need to be considered when the above model and algorithm are deployed in the actual scene.

6. CONCLUSION
This paper gives a comprehensive overview of the relevant technologies and theories in the process of data mining, expounds the application of data mining in the fault detection of mechanical equipment, and points out its shortcomings and development trends. In short, with the advent of the information age, all walks of life have developed rapidly, and data mining has naturally become a powerful application tool and an important auxiliary tool for decision support. Manufacturing industry should keep pace with The Times so that data mining can play a real role in the manufacturing industry.

ACKNOWLEDGMENTS
The research is supported by the National Key R&D Program of China under Grant No. 2017YFE0118700, the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 734599, Shanghai Polytechnic University-Key Discipline Construction (Mechanical Engineering, XXKZD1603) and Graduate program fund of Shanghai Polytechnic University (EGD19YJ0036).

REFERENCES
[1] Tao F, Qi Q. New IT driven service-oriented smart manufacturing: framework and characteristics. IEEE Trans Syst Man Cybern Syst 2017, http://dx.doi.org/10.1109/TSMC.2017.2723764.
[2] Yin S, Kaynak O. Big data for modern industry: challenges and trends [point ofview]. Proc IEEE 2015;103(2):143–6. Another reference
[3] Harding J A, Shahbaz M, Srinivas, et al. Data Mining in Manufacturing: A Review[M]. VDM Verlag Dr. Müller, 2006:969-976.

[4] Wang K, Tong S, Eynard B, et al. Review on Application of Data Mining in Product Design and Manufacturing[C]. International Conference on Fuzzy Systems and Knowledge Discovery, 2007:613-618.

[5] Hand D J. Data Mining[J]. Encyclopedia of Environmetrics, 2006, 2.

[6] Yu C H. Exploratory data analysis[J]. Methods, 1977, 2: 131-160.

[7] Seber G A F, Lee A J. Linear regression analysis[M]. John Wiley & Sons, 2012.

[8] Aiken L S, West S G, Reno R R. Multiple regression: Testing and interpreting interactions[M]. Sage, 1991.

[9] Hansen L K, Salamon P. Neural network ensembles[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1990 (10): 993-1001.

[10] Keller J M, Gray M R, Givens J A. A fuzzy k-nearest neighbor algorithm[J]. IEEE

[11] Hosmer Jr D W, Lemeshow S, Sturdivant R X. Applied logistic regression[M]. John Wiley & Sons, 2013.

[12] Munro D D. Forest growth models-a prognosis[M]/Growth models for tree and stand simulation. Research Note 30. Department of Forest Yield Research, Royal College of Forestry, Stockholm, 1974, 30: 7-21.

[13] Kass R E, Raftery A E. Bayes factors[J]. Journal of the american statistical association, 1995, 90(430): 773-795.

[14] Zhang T, Ramakrishnan R, Livny M. BIRCH: A new data clustering algorithm and its applications[J]. Data Mining and Knowledge Discovery, 1997, 1(2): 141-182.

[15] Inokuchi A, Washio T, Motoda H. An apriori-based algorithm for mining frequent substructures from graph data[C]/European conference on principles of data mining and knowledge discovery. Springer, Berlin, Heidelberg, 2000: 13-23.

[16] Freitas A A. A genetic programming framework for two data mining tasks: classification and generalized rule induction[C]/Genetic Programming 1997: Proc 2nd Annual Conf. Morgan Kaufmann, 1997: 96-101.

[17] Zheng X, Levine D, Shen J, et al. A high-performance computing toolset for relatedness and principal component analysis of SNP data[J]. Bioinformatics, 2012, 28(24): 3326-3328.

[18] Zhang Hongjin. Fault diagnosis of hydraulic System of roadheader based on Fuzzy Neural network [D]. Nanjing University of Science and Technology. 2013.

[19] Xin Xuming. Data Mining System based on process Monitoring and its application [D]. Shanghai: Donghua University, 2004.

[20] Jiao Peng Sha. Research on intelligent device Fault diagnosis based on matrix weighted association rules [D]. Hebei: Yanshan University. 2012.

[21] Hong S J, Lim W Y, Cheong T, et al. Fault Detection and Classification in Plasma Etch Equipment for Semiconductor Manufacturing $e$-Diagnostics[J]. IEEE Transactions on Semiconductor Manufacturing, 2012, 25 (1): 83-93.

[22] Wang. Application of Decision Tree algorithm in mechanical equipment fault diagnosis system[D]. Wuhan: Huazhong University of Science and Technology, 2013.

[23] Xue Yong Shu. Application of vibration fault diagnosis of rotating equipment based on data Mining[D]. Jilin: Jilin University. 2013.

[24] Wang S, Sun X, Li C. Wind turbine gearbox fault diagnosis method based on riem annian manifold [J]. Mathematical Problems in Engineering, 2014, 16(4):835-892.

[25] Jiao bin, Xu Zhixiang. Parameters optimization of LSSVM and application in fault diagnosis of wind power gearbox [J]. Control Enggineering of China, 2012, 19(4):681-686.

[26] Shen B, Liang X, Ouyang Y, et al. Stepdeep: a novel spatial-temporal mobility event prediction framework based on deep neural network[C]/Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 2018: 724-733.
[27] Arora N, Cook J, Kumar R, et al. Hard to Park?: Estimating Parking Difficulty at Scale[C]/Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 2019: 2296-2304