Analysis and application of manufacturing data driven by digital twins

Xiyu Gao¹, Peng Liu¹, ², Qixun Zhang¹, ², ³, Dawei Gao¹ and Xin Huang¹

¹ School of Mechanical and Aerospace Engineering, Jilin University, Changchun, Jilin 130025, China
² Chongqing Research Institute, Jilin University, Chongqing 401100, China
³ E-mail: qixunzhang@jlu.edu.cn

Abstract. This article introduces the importance of data processing to the production process and the application of data in the digital twin workshop. Then the overall process of data processing is introduced. First, several typical data acquisition methods are listed, feed the collected data to the digital twin model, and then extract the model data for integration and convert it into the same XML text format. Perform data cleaning and data analysis on it, and then use the DBN in deep learning to extract and classify the data to complete the real-time matching and transmission of production information to employees. Finally, the future development direction of the data analysis will be briefly summarized.

1. Introduction

With the continuous development of big data, Internet of Things, cloud computing and other technologies, data has become a key factor throughout the entire intelligent manufacturing process [1]. Compared with the traditional manufacturing production model, the core goal of intelligent manufacturing is to monitor the production process of the product with accurate process status tracking and complete real-time data acquisition, thereby realizing the interconnection between the physical world and the information world. Traditional digital workshops mostly only solve the problem of workshop layout planning and production line optimization [2-3]. However, the emergence of digital twin technology provides an effective solution for realizing the integration of the physical world and the cyber workshop, as well as guiding the process of product production, task dynamic scheduling, and real-time equipment monitoring [4]. Therefore, how to achieve a unified standard data collection and processing plan for multi-source heterogeneous data in the workshop is the main problem to solve the visual monitoring of the workshop on-site status. The digital twin's application model as shown in Figure 1.

Aiming at the data application of the digital twin workshop, Chen proposed an aircraft digital twin assembly workshop architecture, research on key technologies such as real-time perception and collection of physical assembly workshop data, virtual assembly workshop modeling and simulation operation technology, data-driven assembly workshop production control [5]. Cao proposed a real-time data acquisition and visual monitoring method for discrete manufacturing workshops based on radio frequency identification (RFID), and designed the functional architecture of the workshop data acquisition and monitoring system based on RFID [6]. Guo proposed a data collection and transmission method of the workshop production data of the intelligent manufacturing terminal according to the data transmission needs of the multi-interface equipment on the smart workshop site.
And based on the industrial control PLC of the workshop production line, the OPC unified architecture was used to complete the data collection work [7].

All in all, the information in the production process of the workshop presents multi-source and massive characteristics. However, due to the lack of effective processing solutions for real-time multi-source information, there will be problems such as missing data, inconsistent format, and logic errors when processing multi-source information. This paper aims to provide a multi-source heterogeneous data processing process to facilitate the timely feedback of data to the corresponding employees and complete the effect of visual monitoring of the on-site status of the workshop.

![Figure 1. Digital twin's application model to data.](image)

2. Data processing architecture

Through the analysis of relevant literature and the on-site investigation of the workshop, this paper proposes a corresponding data processing flow for the digital twin workshop. First, perform real-time collection of heterogeneous multi-source data in the workshop, collect relevant data of manufacturing equipment in real time through detection equipment such as sensors, PLC, etc. Transfer the collected multi-source heterogeneous data to the cloud storage server for sorting and storage, then build the corresponding digital twin model based on the collected data, integrate all the status of manufacturing equipment and reflect it in real time through the model, extract the data in the digital twin model and perform data preprocessing, namely data integration and data cleaning. Then use the DBN intelligent...
decision method in deep learning to analyze and process the data to provide a data basis for production task dynamic scheduling, manufacturing system analysis, fault diagnosis and other services. Match the classified data according to the characteristics of employees in different positions, and appropriately release all kinds of information according to the urgency and importance of the information, send them to the employee's client and APP, enable managers to grasp the status of equipment and positions in real time and enable employees to clarify the current tasks. Realize effective control of the manufacturing process and dynamic scheduling of production tasks, making the production process more efficient, high-quality and stable. The research technology roadmap of this paper is shown in Figure 2.

Figure 2. Technology roadmap of this paper.

3. Methodology

3.1. Data collection
Data collection is the basis of multi-source heterogeneous data processing. Only by realizing accurate and real-time collection of a large amount of raw data generated in the production process and transmitting it to the data storage management platform, can the production equipment, product quality and work scheduling be monitored and managed, thereby helping the management department to make more efficient and precise decisions. In a general digital twin workshop, there are three main types of data that need to be collected: (1) Process data. Mainly take product materials as the tracking object, through data collection, the positioning of product materials can be completed, and the process characteristics of assembly line production can be intuitively characterized. (2) Equipment production process data. It mainly includes movement position data such as the position and rotation angle of the robot and other equipment in the production station of the assembly line. (3) Equipment operating status data. The data is mainly for monitoring the working status of the production line station, which can intuitively reflect the working condition of the workshop [8].
According to the collection frequency, sampling accuracy, and data collection volume of multi-source manufacturing data, the selection and configuration of heterogeneous sensing equipment are carried out. For data in discrete manufacturing, RFID technology is mainly used to collect data on raw materials, equipment, and product information in the production workshop [9]. For the data in assembly line production, the data is collected mainly by sensors and upper computer. Chen proposed a method of data collection using auxiliary control systems and control devices such as distributed control system (DCS) and programmable logic controller (PLC) [10]. In process production, computer-based data acquisition systems also include supervisory control and data acquisition (SCADA) systems. Among them, PLC is mainly used for temperature measurement and control in the production site, DCS is mainly used for data collection in the production site that requires high measurement and control accuracy and speed, and SCADA combines the on-site measurement and control function of PLC and the networking communication ability of DCS, which can control dispersion point, so as to achieve coverage of a wide range of production sites [11].

3.2. Intelligent modeling
The above-mentioned processing equipment equipped with smart instruments such as sensors and RFID will form an interconnected sensor network through programmable technology and networked layout. Use industrial computers and sensor networks to establish a complete mapping relationship between manufacturing service status and perception events, in order to construct a real-time multi-source information active perception model, that is, a digital twin virtual model. Then transform, filter, analyze, and process sensor signals to achieve real-time and accurate control of multi-source equipment production data. Construct the perception neural network of the digital twin intelligent system, and extract data from the model to facilitate seamless and real-time two-way interconnection between the production process and the digital platform.

The virtual model mainly includes geometric model, physical model, behavior model and rule model [12]. (1) The geometric model describes the three-dimensional model of geometric parameters (such as shape, size, position, etc.) and relationships (such as assembly relationship). The geometric model can be obtained by rendering the model through 3Dmax, CATIA, and other software. (2) The physical model adds the information of the physical attributes, constraints and characteristics of the intelligent instrument on the basis of the geometric model. Import the geometric model into the NX software and use the MCD module to add a physical field to the model to obtain a physical model, and complete the preliminary visualization of the workshop production line. (3) The behavior model describes the real-time response and behavior of smart instruments at different spatial scales under different time scales under the external environment and interference, as well as the internal operating mechanism, such as evolutionary behavior and dynamic functional behavior, etc. That is, after data collection, the data is transmitted to the digital model to establish communication to complete the movement state matching. (4) The rule model includes regular rules based on historical associated data, experience based on tacit knowledge, and related field standards and guidelines. That is, the information fed back to the digital twin model after data is integrated and analyzed and deep learning.

3.3. Data integration
In view of the different data structures and inconsistent description methods of the upper and lower systems, the data cannot be understood between the systems, the XML working group of the World Batch Forum has developed the enterprise manufacturing markup language B2MML to provide manufacturers with free ISA-95 enterprise XML format applications to make them comply with control system integration standards [13]. The ISA-95 standard provides modules and terminology to standardize the multi-source and scattered field data collected on the upper-level heterogeneous sensors, and uses XML Schema to describe the resources and data flows defined by the standard to define the content and format of data exchanged between enterprise management systems. The data integration and conversion model is shown in Figure 3.
3.4. Data cleaning
Accurate and reliable data is the prerequisite for effective data analysis and data mining [14]. In the production process, due to the many characteristics of multi-source heterogeneous data sources, the types of data collected by sensors are diverse, and the collected raw data often has problems such as missing data, logical errors, inconsistent format and data duplication, the quality of the collected data is difficult to guarantee. The purpose of data cleaning is to detect the "dirty data" in the data, and improve the quality of the data by means of data screening and data repair.

For missing data, it is necessary to fill in manually in most cases, and in some cases, the missing values can be processed by statistical learning methods. For error data, it is necessary to use statistical analysis methods to identify possible error values, and then to clear the wrong data to achieve the purpose of data cleaning. For inconsistent data, potential errors in data can be detected based on the consistency between associated data and repaired to complete the cleaning of data from multiple data sources. Cleaning duplicate data is mainly divided into three steps: pre-cleaning, detection rule learning and result clean [15]. The preprocessing stage is mainly to process different XML documents, extract the root elements required by the query and perform the following processing to facilitate the learning of the subsequent repeated recognition rules and the final repeated element recognition. It mainly includes the following tasks: (1) perform spell check (2) unify the case of character data (3) deal with redundant symbols and blanks (4) establish a synonym table of tags to facilitate the learning of conversion rules in the following. Detection rule learning is mainly for specific applications, learning conversion rules and matching rules for repetitive element recognition based on limited training samples, which is to identify XML repetitive elements key. The role of the cleaning phase is to use the learned conversion rules and matching rules to establish a hash table for XML data to identify repeated elements.

3.5. Data analysis
Data analysis is the process of letting the computer classify the text set of a given unknown category according to the established classifier [16]. A complete text classification process mainly includes the
following parts: first, text preprocessing, which represents the text into a form that is easy for computer processing. The second is the text vector representation. Thirdly, learn modeling based on the training set (with class labels) to construct a classifier. Finally, use the test set to test the performance of the established classifier, and continuously feedback and learn to improve the performance of the classifier until the predetermined goal is reached.

Deep learning is to construct a machine learning model with multiple hidden layers and combine low-level features to form more abstract high-level features to represent attribute categories to discover distributed features of data. Deep learning has a very powerful self-learning ability to mine features, can obtain expressions and patterns that are closest to the essence of the data, and can greatly improve the performance of prediction and classification. In 2006, Geoffrey Hinton proposed deep belief network (DBN) [17]. It is a probability generation model. By training the weights between its neurons, the entire network can generate training data according to the maximum probability, thereby achieving feature recognition and data classification. The deep belief network is composed of multi-layer neural networks, and these structures are all non-linearly mapped, which makes it possible to complete the approximation of complex functions well, thereby achieving fast training [18].

3.6. Data application

After processing the workshop data completely, it can be mainly applied to the following two aspects:

(1) After obtaining the complete workshop processing data, give feedback to the twin model, continuously update the twin model’s operating data and fault data, improve the high fidelity of the digital model, monitor the status of the workshop equipment and evaluate the health, and establish the equipment fault prognostics model to complete fault prognostics.

(2) After classifying the processing data, set the corresponding characteristics of the employees in each position, match the characteristics of the data extracted with the characteristics of the employees, and transmit the information belonging to each employee category through the client or mobile APP. Including the operating status of equipment, product quality information, processing progress, etc., it provides help for management personnel's information control and task dynamic scheduling.

4. Discussion

With the rapid development of the Industrial Internet of Things, there are more data sources and more diversified data structures. At the same time, the information system in the production process has higher requirements for the real-time and accuracy of data processing, which brings the processing of multi-source heterogeneous data. In the future, more abundant data acquisition methods, faster data storage units, and more accurate and efficient data processing and analysis methods are needed to better promote the development of digital twins and smart instruments. Moreover, the deep learning algorithm will degrade step by step under a series of process changes such as equipment update, personnel replacement, and production process changes. How to effectively maintain and improve the modeling accuracy and performance of deep learning, iterative optimization of algorithms, real-time feedback updates of digital twin models, future trend prediction, demand analysis and risk assessment, and other intelligent decision-making service needs are all difficult points in the field of digital twin information research in the future.

5. Conclusions

This paper discusses the importance of data to the production process in the digital twin workshop, then introduces the overall process of multi-source heterogeneous data processing, including data collection, digital twin modeling, data extraction, integration and conversion, data cleaning, data analysis and data application. The complicated data can be classified and sent to the corresponding managers at all levels and employees in each position to remind their respective tasks. Through the processed data feedback digital model, continuous iterative optimization, to achieve remote monitoring and fault prognostics. Compared with traditional production management methods, it can improve the speed and efficiency of management and manage the workshop visually.
Acknowledgements
This research was supported by the National Natural Science Foundation of Jilin Province (20200201231JC) and "Thirteenth Five-Year Plan" Science and Technology Project of the Education Department of Jilin Province (JJKH20200956KJ).

References
[1] Yu M, Zhu M, Chen G and Li J S 2016 A cyber-physical architecture for industry 4.0-based power equipments detection system International Conference on Condition Monitoring and Diagnosis Xi'an pp 782-785
[2] Zhao R Y, Zhang H and Fan L Q 2004 The research about relationship of digital factory and virtual manufacturing Computer Integrated Manufacturing Systems S1 46-50
[3] Ning R X, Liu J H and Tang C T 2006 Modeling and simulation technology in digital manufacturing Chinese Journal of Mechanical Engineering 42 132-137
[4] Himmler F and Amberg M 2014 Data integration framework for heterogeneous system landscapes within the digital factory domain Procedia Engineering 69 1138-43
[5] Chen Z, Ding X and Tang J J 2018 Digital twin-based production management and control mode for aircraft assembly shop-floor Aeronautical Manufacturing Technology 61 46-50+58
[6] Cao W, Jiang P Y and Jiang K Y 2017 Radio frequency identification-based real-time data collecting and visual monitoring for discrete manufacturing workshop Computer Integrated Manufacturing Systems 23 273-284
[7] Guo L, Chen X Y and Zhang Y L 2019 Data collection and transmission method for workshop production in intelligent manufacturing terminal Machinery & Electronics 37 21-24
[8] Lin C Y, Zhou J W and Shi J C 2020 Design of data acquisition system for digital twin Electronics Quality 10 71-75
[9] Chen S C, Cui C Y, Zhang H and Ma G 2020 Summary of multi-source heterogeneous data processing methods in manufacturing production process Big Data Research 6 55-81
[10] Chen K S 2017 Research on data acquisition technology in manufacturing industry Journal of Kaifeng University 31 93-96
[11] Liu S F, Chen X Y and Zhang Y H 2017 Application of Honeywell SCADA system in urban gas pipeline network Jiangsu Science and Technology Information 12 56-57+60
[12] Tao F, Cheng J F and Qi Q L 2018 Digital twin-driven product design, manufacturing and service with big data The International Journal of Advanced Manufacturing Technology 94 3563-76
[13] Yue L, Niu P F and Wang Y 2019 Guidelines for defining user requirement specifications (URS) of manufacturing execution system (MES) based on ISA-95 standard Journal of Physics: Conference Series 1168
[14] Ainhoa A, Marta P and Ivan B 2017 Massive data load on distributed database systems over HBase 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (Madrid: IEEE) pp 776-779
[15] Melanie W and Felix N 2004 Detecting duplicate objects in XML documents 2004 Proc. of the 2004 Int. workshop on Information quality in information systems 10-19
[16] Hinton G E and Salakhutdinov R R 2006 Reducing the dimensionality of data with neural networks Science 313 504-507
[17] Hinton G E, Osindero S and Teh Y 2006 A fast learning algorithm for deep belief nets neural computation Neural computation 18 1527-1554
[18] Wang Y L, Pan Z F, Yuan X F, Yang C H and Gui W H 2020 A novel deep learning based fault diagnosis approach for chemical process with extended deep belief network ISA Transactions 96 457-467