Repairing of Missing Data in Scheduling Based on Generative Adversarial Network

Jing Chen

Hangzhou Dianzi University

Ying Song

Hangzhou Dianzi University

guodong yi (✉ yiguodong234@hotmail.com)

Zhejiang University

Shuyou Zhang

Zhejiang University

Wanghui Bu

Tongji University

Research Article

Keywords: Generative adversarial networks, Data repair, Scheduling problem, Smart manufacturing, Data imputation

Posted Date: March 7th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1256263/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

In the manufacturing process of modern smart factories, when product family design and variant design are carried out on the basis of existing products, local design changes will generate new technological processes. Process changes lead to partial lack of scheduling data, which in turn affects the efficiency of workshop scheduling in the manufacturing process. Most of the research results related to scheduling problems are focused on the optimization of various scheduling objectives, without considering the problem of missing scheduling data. Therefore, this paper proposes a data processing method that combines missing data removal and data expansion based on processing order for optimal training of GAN to solve the problem of insufficient training data when the scheduling data set is small. In this paper, We use GAN to generate local data, and combine the missing vector of the data and the hint vector to strengthen the training process, complete the repair of the missing data, and then further study the effectiveness of the repair. Experimental results show that this method of using historical process data to repair scheduling data can effectively repair the problem of partial missing data in the same product family and improve the efficiency of workshop production scheduling.

1. Introduction

The job shop scheduling problem(JSP) has applications in many modern manufacturing industries, such as engine assembly, semiconductor manufacturing, and automotive parts processing. In real workshop manufacturing, given multiple sets of workpieces to be processed, each set of workpieces has tasks that need to be processed by the machine in a specific order, which is a series of complete processing steps. Under the conditions of occupancy restrictions and resource restrictions, many previous works, such as [1–3] use Current production data (such as environmental information, operating requirements, etc.) to find a series of decisions to allocate resources to optimize one or more goals (for example, total processing time, average processing time, and cost). The premise of these studies is that the scheduling data is complete and accurate. However, due to product family design and variant design on the basis of existing products, local design changes will generate new technological processes, resulting in partial invalidity of scheduling data in the new technological process. Or due to incomplete records, loss, damage, etc., the lack of scheduling data occurs in the manufacturing process. As described in [4], Intel's engineering and business operations heavily rely on accurate manufacturing data. Due to the inaccurate engineering analysis, the key data lost during the manufacturing/sorting/assembly/testing manufacturing process brings Intel the risk of shipping substandard products to the end customer, no single automated system can eliminate the risk of data loss, which adds to the complexity of the problem. The loss of data not only causes the processing target task to be unable to be completed as scheduled, but also increases the unqualified rate of processed products, thereby increasing the cost. In a series of processing steps, the number of processes of the workpiece means that there are more error accumulation links, and the lack of any processing time will make the workpiece waste. Since products often contain multiple processing steps, a very low rate of missing data can lead to a high rate of scrap and rework. Moreover, the job shop scheduling data set is generally small in scale, with a small number of

workpieces and machines. The procedures and processing times of different types of processed parts are very different, and the location of data missing is random. Once the data is missing, it is difficult to find references from other datasets.

In the case of invalid or missing scheduling data, how to achieve the expected scheduling tasks to the greatest extent and reduce losses is a problem that smart manufacturing factories need to solve in the context of the fourth industrial revolution (I4.0). How to use contemporary artificial intelligence technology to enable intelligent processing workshops to have the function of responding to changes in processing requirements and emergencies in real-time under the premise of meeting scheduling requirements, thereby improving the flexibility and efficiency of the manufacturing industry. As Parente et al. [5] put forward, modern society has increasingly higher requirements for smart devices, requiring them to be able to deal with different unexpected problems. In the face of uncertain and incomplete scheduling data, smart devices can provide as effective manual data as possible.

2. Related Work

In reality, data is missing in many fields, and many solutions have been proposed accordingly. In the building field, the loss of building operation data has a great impact on the building energy system. Zhang [6] uses a framework that combines pattern recognition and integration methods to improve the overall accuracy of data interpolation; Krause et al. [7] studied and compared different missing value interpolation methods for missing data in the network, and showed that in the case of a large amount of missing data, the complex classification model based on the exponential random graph model will produce acceptable deviations; In the field of software development, the loss of workload data usually occurs. Jing et al. [8] proposed a novel low-rank recovery and semi-supervised regression imputation workload data imputation method, Song et al. [9] proposed a class mean interpolation method based on the k-NN hot deck interpolation method; Gebregziabher et al. [10] proposed a multiple imputation method based on latent classes to analyze the missing categorical covariate data in the highly hierarchical data model; In response to missing data in the Internet of Things field, Yan et al. [11] divided missing data into three types, and proposed three new models: context and linear mean missing value interpolation Model, missing value imputation model based on binary search and missing value imputation model based on Gaussian mixture model .

In recent years, with the development of Generative Adversarial Network technology, GAN combined with other deep learning technologies has more and more applications in data processing Yang et al. [12], Wang et al. [13], Yu et al. [14], which has solved many problems. In the field of job shop scheduling, there is no well-established and mellow technology to solve the problem of missing scheduling data. Therefore, this paper uses the technology based on generative adversarial network to study the problem of missing data for job shop scheduling workpiece processing time.

The main contributions of this paper are as follows:

1. This paper proposes a Generative Adversarial Network (GAN) scheduling data generation method based on the process data of the same product family, generates partial data, completes the repair work of processing time in scheduling manufacturing, solves the problem of data loss in small and medium-scale scheduling data sets.
2. Aiming at the small and medium-sized workshop scheduling data set, this paper proposes a data processing method that combines missing data removal and data expansion based on scheduling sequence, which is used for GAN optimization training and solves the problem of insufficient training data.

3. Scheduling Data Description

3.1 Problem definition

In the job shop, there is a group of small-scale processing workpieces, the number of which is n ($N=\{n_1, n_2, \dots, n_{20}\}$), and the number of processing machines is m ($M=\{m_1, m_2, \dots, m_{20}\}$), each workpiece has m processing steps. The machine processing of the same set of workpieces is different, and the processing time of each workpiece on each machine is also different. One of the small parts of the processing time data is missing, which causes the unqualified rate of the workpiece to rise. The purpose of the paper is to generate missing data based on the existing information of scheduling data, and the generated processing data should be proximal as possible to real data.

The overall framework of this paper is shown in Figure 1. The missing data generated from the original data is input to the generative adversarial network after data processing, and the generative adversarial network outputs the repaired data. The specific content is described in the following chapters.

3.2 Data processing

For the scheduling data set with the missing part of the processing time, this paper proposes the following data processing methods.

1. Data expansion: Due to the small scale of the job shop scheduling data set, and the data rules between different data sets are rarely correlated and cannot be referred to each other, this paper proposes a data expansion method based on the scheduling sequence for GAN Training optimization. That is, regardless of whether the workpiece lacks processing time data, each workpiece information is processed through "machine number, processing time, processing serial number". As shown in the data expansion part in Figure 2, where 'M3', 't11' indicates that the first processing machine of the first workpiece is 'M3', and the processing time on the 'M3' machine is 't11'; 'M1' indicates that the second processing machine is M1, 'miss' indicates that the processing time data on the 'M1' machine is missing. After extended processing, 'M3', 't11', '1' indicating that the first processing step is to process t11 time units on the 'M3' machine; 'M1', 'missing', '2' indicating that the second processing step is processed on the 'M1' machine, and the processing time is missing.

2. Missing data removal: In the data-expanded scheduling data set, each processing step is an independent piece of data. Delete the missing processing time steps and use the remaining process data of the same product family to generate manual data. As shown in the green frame in Figure 2.

4. Data Repair

4.1 Scheduling data repair based on Generative Adversarial Networks

The structure of the network is shown in Figure 1. It follows the framework of GAN, the network is divided into two sub-networks, a generator, and a discriminator. The generator generates scheduling data, and the discriminator distinguishes whether the scheduling data is real data or artificial data. In this way, The generator continuously improves the effectiveness of the generated data, and the discriminator continuously improves the discriminative power., as shown by arrows in Figure 1. In an ideal situation, the artificial data generated by the generator is very close to the real scheduling data.

As shown in Fig. 1 and Fig. 2, GAN is composed of two parts: generator G and discriminator D, both G and D are modeled as fully connected neural layer, and optimize parameters with the back propagation algorithm. Generator G generates complete data (processing time for each process) according to the original incomplete data(The workpiece lacks part of the processing time data). The discriminator D generates an estimated matrix to classify the machining time. Judging the machining time, which is true and which is generated by generator G. Using the Mark Matrix and Hint Matrix methods in Yoon et al. [15], after training, generator G and discriminator D has achieved the best performance to each other, so that the generated data is close to the real processing time. Mark Matrix is a matrix consisting of {0,1} elements, defined as M. The role of M is to mark the position of the missing data and calculate loss function, '1' means that the scheduling data of this position is existing, '0' means that the scheduling data of this position is missing. In the generator G, the original data Containing missing data is defined as T, and artificial data is generated according to T. According to M extracts the missing data position, the artificial data is combined with the original scheduling data into complete predictive data.

Hint Matrix is defined as H, the role of H is to prompt the discriminator D to miss data information, thereby increasing the performance of the generator G, enabling the data output by the generator G close to real data. When the data in the H is 0 or 1, it is shown that the value of this position H is equal to M; when the data in the H is another value, indicating that the value of this position M cannot be determined. The output of the discriminator is defined as $\widehat{M}$. Multiple training generator G and the discriminator D generate final data for actual scheduling processing. The final data is defined as $\widehat{T}$.

Objective:

$$\text{Where } \widehat{M} = D(\widehat{T}, H), \widehat{T} = G(T, M)$$

4.2 Analysis Method for effectiveness of Artificial Data

The absence of any processing step will make the workpiece a substandard workpiece. In this paper, we considered that artificial data is effective if the absolute error between artificial data and real data is within the specified time range.

The absolute error is defined as:

$$\delta = |d - \hat{d}|$$

Where d is the real data, and $\hat{d}$ is the artificial data. The value of the error range for judging whether the manual data is valid is related to the processing time range of the original data set. If the processing time range of the original JSP data set is too large, the time range is 10-time units; if it is too small, the time range is 5-time units. (Time unit: take the time unit of the original JSP data set)

5. Experiment

The data set in the experiment comes from two parts: (1) "Adams, Balas and Zawack" and "Yamada and Nakano" in the JSP data set [16], the scales (number of workpieces $\times$ number of machines) are: 10 $\times$ 10, 20 $\times$ 15, 20 $\times$ 20; (2) The case of automobile engine assembly line.

The original data set is complete. Part of the data selected in this paper is set as missing, as shown by arrow in Figure 1. Due to the different categories of the two data sets, set separately: the data missing rate of the processing time of the data set (1) is 5%, and each workpiece in the same group has a certain probability of missing 1~2 processing time, causing the entire workpiece to become unqualified Product; the data missing rate of the processing time of the data set (2) is 10%.

5.1 Error performance analysis of scheduling data repair

This paper uses eight different scheduling data cases in the job shop scheduling data set to test the performance of the model. Each scheduling data case is processed for data normalization, and then the model is run ten times to observe the error between the generated data and the real data. Considering that the data rules in different scheduling data cases are different, the actual performance of different scheduling data cases on the model is not the same. The experimental results conform to this conjecture.

As shown in Figure 3 and Figure 4, the horizontal axis is the order in which the data cases are run on the model, a total of ten times, and the vertical axis is the root mean square error (RMSE). The RMSE performance of different scheduling data cases is different, but all the RMSE tend to be stable, and the average RMSE is below 0.35. As shown in Fig. 3, the average RMSE of YN01 (20 $\times$ 20) is the largest(0.334); the average RMSE of YN04 (20 $\times$ 20) is minimal(0.226). In the case of the same data missing rate, the scale of the JSP data set (number of workpieces $\times$ machine) cannot determine the

performance of the RMSE. For example, in Fig. 2, the scale of abz5 is 10×10 , and its RMSE does not significantly greater than or less than the scale of 20×15 JSP data set(abz7, abz8, abz9).

Table 1
Mean RMSE of the abz series and yn series

Size	Name	Mean	Size	Name	Mean
10×10	abz5	0.312	20×20	yn01	0.334
20×15	abz7	0.265	20×20	yn02	0.267
20×15	abz8	0.329	20×20	yn03	0.273
20×15	abz9	0.3	20×20	yn04	0.226

5.2 Validation of data repairing

Different scheduling data cases have different processing requirements and processing time rules. After data repair, their actual error performance is different, which is confirmed in experiments. Suppose the missing data is invalid, and the absolute error () of the generated artificial data exceeds a certain time range as invalid data. A workpiece has multiple processing data, and invalid data will make the workpiece not meet the processing requirements.

Define P as $P = \text{Number of workpieces containing invalid data} / \text{Number of workpieces}$. When all the missing data of a workpiece are effectively repaired, the value of P will change.

As shown in Fig. 5, Table 2, after the data repair processing, the difference between the two p values of YN04 is the largest; and the difference between the two P values of YN03 is minimized. After data repair processing, the number of effective time data has increased, which proves that the work of this paper has certain practical significance for the problem of missing data in job shop scheduling.

Table 2
Comparison of scheduling data validity after data repair

Time Range	Size	Name	Proportion of invalid data before repair	Proportion of invalid data after repair
10	10×10	abz5	0.060	0.030
5	20×15	abz7	0.046	0.033
10	20×15	abz8	0.046	0.023
5	20×15	abz9	0.053	0.033
5	20×20	yn01	0.045	0.032
5	20×20	yn02	0.0525	0.030
5	20×20	yn03	0.0475	0.032
5	20×20	yn04	0.0425	0.022

5.3 Data repair analysis of engine assembly line

Because the data in 5.1 is mainly used for comparison of scheduling optimization results. To analyze the effectiveness of data repair in the actual manufacturing process, we took the actual scheduling data in the automobile engine assembly line as the object and conducted experiments on data repair and effectiveness analysis.

The automobile engine assembly line contains a total of 291 processes, and the cycle time constraint is 66 seconds. There is a sequence of constraints between each process, part of which is shown in Figure 6.

Each process has its characteristics, such as automatic type, manual type; type of orientation; processing time, etc. As shown in Figure 6. The first column "A" represents that the process is automatic, and as shown in the fourth column, the cycle time of each automatic process is 66 seconds, and it does not need to be combined with other processes. The automatic process only needs to meet the constraints of the sequence relationship and arrange it into the corresponding station; "M" means that the process is a manual process, and the cycle time requirement of the manual station is less than or equal to 66 seconds. The cycle time of each manual process is different, and it needs to be combined and sorted according to the sequence constraints and the cycle time of the station, and finally arranged into the corresponding station. The second column is the actual name of the process, which is also the actual operation process. The third column represents the orientation type of each process. There are 4 orientation types for all 291 processes, namely "FRT", "INT", "REAR" and "EXH". If there is a change in the orientation type between two adjacent manual stations during the process combination, the latter process needs to add 3 seconds. If the change is from a manual station to an automatic station, or there is a change in the orientation type from an automatic station to a manual station, then 3 seconds need to be added to the manual workstation. The fourth column is the respective cycle time of each process. Among them, the cycle time of the automatic process is 66 seconds, and the cycle time of each manual process

is different according to the actual working conditions. ">" or ">=" represents the processing order constraint. Among them, ">" means that the process corresponding to the row must be completed after the process corresponding to the column is completed; and ">=" means that the process corresponding to the column must be executed immediately after the process corresponding to the column is completed Process.

The process combination optimization problem of the assembly line requires that each process be reasonably arranged in the workstation under the condition of satisfying the cycle constraint and the sequence relationship between the processes. The goal is to minimize idle time at the workstations.

The number of workstations and the distribution of processes has been determined. The processing time of some workstations is missing due to process changes, inaccurate engineering analysis, or incomplete records. Our method repairs missing data based on existing data, with the goal of artificial data as close to real data as possible.

In order to further improve the accuracy of manual data, a new data classification processing method is added in the data processing step. According to the process type (automatic type (A) or manual type (M)) in the workstation, the data is divided into two categories to achieve more accurate data generation effects. As shown in Figure 7.

The manual data performance of such case data sets is relatively stable. Since the time of the automatic process station is fixed, the missing data can be accurately generated. The RMSE of the missing data of the automatic station is 0, and the RMSE of the missing data of the manual station is 0.183. If the data is not classified, the RMSE performance of the missing data is 0.269. From the experimental results, it can be seen that processing the data separately according to the process type of the workstation will achieve better results.

Using the method in this paper, the missing scheduling data repairs in the automobile engine assembly line are all effective repairs. Then, we further analyze the relative error between artificial data and real data, which is defined as:

$$\delta 1 = \left| \frac{d - \hat{d}}{d} \right|$$

Where d is the real data, and $\hat{d}$ is the artificial data.

As shown in Figure 8, the vertical axis is d , and the horizontal axis is the missing data. For example, A1 is the first data missing in the automatic process station, and M1 is the first data missing in the manual process station. Missing data means that no repairs are made to the missing data; the way to randomly generate the time is to take a random number between the maximum time data and the minimum time data. Obviously, the repair effect of the method in this paper is better.

6. Conclusion

This paper proposes a data processing method that combines missing data removal and data expansion based on scheduling sequence, and studies the problem of data missing in small-scale job shop scheduling based on the technology of generative adversarial network and further explore whether the missing data repair work is effective for JSP. In actual factory manufacturing, the missing data is inevitable. The experiment proves that after repair, the number of invalid data in JSP data is significantly reduced, which has a certain actual significance for workshop processing.

Since there are many types of JSP datasets, these data sets are different in the number of workpieces, the number of machines, the processing time, the type of procedures, the number of stations, etc. How to find the data feature of the JSP dataset, make the data repair performance better, and find a unified method for all types of JSP datasets, which is the future that needs to be considered in the future.

Declarations

Acknowledgments

This work is supported in part by the National Key Research and Development Program of China under Grant 2018YFB1701300, Zhejiang Provincial Natural Science Foundation of China under Grant LY21F020014, LY17F020026 National Science Foundation of China under Grant 61703127.

Authors' Contributions

JC and YS conceived the idea and built the methodology. JC, YS, and GY wrote the original draft. SZ and WB supported with their extensive experience and gave advices on the manuscript. All authors approved the final manuscript.

Funding

Part of this work was supported by the National Key Research and Development Program of China, grant number 2018YFB1701300 National Science Foundation of China under Grant, grant number 51475331, 61703127.

Availability of Data and Materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing Interests

The authors declare that they have no competing interests.

References

1. C-C Lin, D-J Deng, Y-L Chih, et al. Smart Manufacturing Scheduling With Edge Computing Using Multiclass Deep Q Network. *IEEE Transactions on Industrial Informatics*, 2019, 15(7): 4276-84.
2. D Tang , M Dai. Energy-efficient approach to minimizing the energy consumption in an extended job-shop scheduling problem. *Chinese Journal of Mechanical Engineering*, 2015, 28(5): 1048-55.
3. N Zhao, S Chen, DU Y. Emergency local searching approach for job shop scheduling. *Chinese Journal of Mechanical Engineering*, 2013, 26(5): 918-27.
4. L Y Lin. Enable Robustness of Data System for Intel Assembly/Test Manufacturing Factories with High Degree of Data Integrity. *International Electronic Manufacturing Technology*, 2006.
5. M Parente, G Figueira, P Amorim, et al. Production scheduling in the context of Industry 4.0: review and trends. *International Journal of Production Research*, 2020, 58(17): 5401-31.
6. L Zhang. A Pattern-Recognition-Based Ensemble Data Imputation Framework for Sensors from Building Energy Systems. *Sensors (Basel)*, 2020, 20(20).
7. R W Krause, M Huisman, C Steglich, et al. Missing Network Data A Comparison of Different Imputation Methods. *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 2018: 159-63.
8. X-Y Jing, F Qi, F Wu, et al. Missing data imputation based on low-rank recovery and semi-supervised regression for software effort estimation. *Proceedings of the 38th International Conference on Software Engineering*. 2016: 607-18.
9. Q Song, M Shepperd. A new imputation method for small software project data sets. *Journal of Systems and Software*, 2007, 80(1): 51-62.
10. M Gebregziabher, S M Desantis. Latent class based multiple imputation approach for missing categorical data. *Journal of Statistical Planning and Inference*, 2010, 140(11): 3252-62.
11. X Yan , W Xiong, L Hu, et al. Missing Value Imputation Based on Gaussian Mixture Model for the Internet of Things. *Mathematical Problems in Engineering*, 2015, 2015(1-8).
12. M Yang, J Liu, L Chen, et al. An Advanced Deep Generative Framework for Temporal Link Prediction in Dynamic Networks. *IEEE Trans Cybern*, 2020, 50(12): 4946-57.
13. Q Wang , Y Ji, Y Hao, et al. GRL: Knowledge graph completion with GAN-based reinforcement learning. *Knowledge-Based Systems*, 2020, 209.
14. J J Q Yu , J Gu. Real-Time Traffic Speed Estimation With Graph Convolutional Generative Autoencoder. *IEEE Transactions on Intelligent Transportation Systems*, 2019, 20(10): 3940-51.
15. J Yoon ,J Jordon,M Schaar. GAIN: Missing Data Imputation using Generative Adversarial Nets[J]. 2018.
16. Jobshop Instance, 2015. [Online]. Available: <http://jobshop.jjvh.nl/index.php>. Accessed on: 2021.

Figures

Figure 1

Overall framework

Figure 2

Data processing and repair

Figure 3

RMSE performance of the abz series

Figure 4

RMSE performance of the yn series

Figure 5

P value of scheduling data

Figure 6

Processes of automobile engine assembly line

Figure 7

Data classification of automobile engine assembly line

Figure 8

The performance of the relative error of missing data

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [automobileengineassemblyline.xlsx](#)