Quality Prediction of Discrete Manufacturing Process Based on CGAN&Catboost Hybrid Model

Peng Xu\textsuperscript{1}, Zhichen Ren\textsuperscript{2}, Yili Shen\textsuperscript{2}, Wendi Yu\textsuperscript{2}, Lianghua He\textsuperscript{2,*}

\textsuperscript{1}School of Computer Science & Engineering, Chongqing Jiaotong University, Chongqing, China, 400074

\textsuperscript{2}Department of Computer Science and Technology, Tongji University, Shanghai, China, 201804

\textsuperscript{*}xupeng@mails.cqjtu.edu.cn

Abstract: Accurate prediction of product quality can be used to control the online production process of workpieces, an appropriate prediction model can effectively improve the yield of industrial production. In order to obtain missing data, the classic method is to regress them one by one. However, the traditional regression prediction method destroys the correlation of the data and cannot use the quality inspection index efficiently. This paper presents an improved hybrid model. The essence lies in a novel easily pluggable CGAN module that refines the previous step of regression data. Compared with the regression method, the CGAN module can learn the joint distribution of data. The experiment of real data result shows that our method has high accuracy compared with the existing ones.

Keywords: CGAN Module, Hybrid Model, Quality Prediction, Industrial manufacturing

1. Introduction

Along with the development of industrialization in human society, a great deal of discrete manufacture data is increased rapidly. At present, practitioners in the manufacture field rely on observations from process parameters to product parameters and to decide an optimization strategy for product quality levels. To the best of our knowledge, the data-driven models have been widely used in many industrial fields such as predicting the industrial production process through machine learning algorithms and further optimizing the process through controlling parameters, ultimately improving the efficiency of
industrial production [1-5], diagnosing industrial process faults through data analysis [6,7], tremor prediction in industrial processing using big data [8].

In industrial production, product parameters can only be obtained after the process parameters are determined. This type of information can be used in model training but it cannot be used in model testing called privilege information (PI). For the acquisition of high-quality product parameters from industrial data, an intuitive idea is to train a series of regression models for each product parameter feature, and the product parameter classification effect obtained by regression is improved. The bottleneck lies in that regressing each product feature separately requires a prior hypothesis that the product parameters are independent of each other, but in fact, there must be a correlation among various parameters of the product. The model breaks up this correlation. If the mapping relationship between historical data process parameters and product parameters can be learned through a model, the classification effect of process parameters on product quality levels can be improved.

To this end, we propose a hybrid model that utilizes a CGAN (Conditional Generative Adversarial Network) module, which is readily pluggable into the existing classification model, to generate high-quality product parameter distributions and be classified by Catboost. Firstly, using the CGAN module to generate high-quality data. Then cascade the generated data by CGAN and raw data. Finally, the cascade data is classified by using the Catboost algorithm. Experiments show that hybrid models have a certain improvement in accuracy compared with the original model.

2. Data introduction and preprocessing

The collected dataset largely determines the quality of the data-driven model. In this study, we take a typical production process as a dataset, which is provided by CCF BDCI2019*. The dataset is collected by the factory and has been desensitized, which contains a series of process parameters and the quality data of the workpiece produced by the corresponding process parameters.

The original data contains 18934 samples. In order to evaluate the model, the dataset is randomly divided into two parts: Taking 15934 samples as a training set and 3000 samples as a testing set.

The samples contain the following information.

A: Process parameters (such as equipment processing parameters)
B: Quality data of the workpiece
C: Quality inspection index that the workpiece meets(label)

There are ten process parameters (A, Parameter1-10), ten quality parameters (B, Attribute1-10), and quality inspection indicators are divided into four categories(C). The privilege information(PI) in this paper is B (quality data of the workpiece).

Special Instructions: The training data includes process parameters, quality parameters, and their corresponding label (A, B, C). The testing data only includes process parameters and corresponding quality inspection indexes (A, C).

2.1 Data preprocessing

Firstly, the correlation degree of the data was analyzed and it was found that Parameter1-4, Attribute1-3 are not strongly correlated with the rest of the features. Meanwhile, the distribution on the label shows no differences. These seven features are of low importance in the model training process, so the Parameter1-4 and Attribute1-3 features are deleted. Secondly, the combination feature
Parameter11 is constructed by analyzing the data correlation as Parameter5+1.05* Parameter6*, which was used as one of the features of the final model. In addition, Parameter9 was deleted because it misses a lot in the test data. Finally, we perform Box-cox transformation on the selected features one by one to improve the normality, symmetry, and equality of variance of the data.

The features are used in the hybrid model include non-privileged information Parameter (5, 6, 7, 8, 10, 11), privileged information Attribute (4, 5, 6, 7, 8, 9, 10), and label.

3. Algorithm introduction

3.1 Generative Adversarial Network (GAN)

Generative Adversarial Networks (GAN) is one of the most promising methods for unsupervised learning on complex distributions in recent years. Because the joint distribution of data can be obtained through training, GAN was widely used for data generation [9]. At present, the main purpose of GAN is to generate discrete data and it can increase the number of available data in training. Since the mapping of random noise to training samples is learned, the data after expansion performs better. The purpose of using GAN in this article is to use the GAN network to achieve the mapping function from low-quality data (non-PI) to high-quality data (PI) by learning the joint distribution.

The generative adversarial network GAN can learn the data distribution through game training between the generator and the discriminator and then generate new samples. But the disadvantage of GAN is that the generated images are random and cannot control which category the generated images belong to. CGAN [10] is the progress made on the basis of GAN. It implements a conditional generation model by adding additional conditional information to the original Generator and Discriminator.

3.2 Catboost

GBDT is a very popular and effective algorithm model in machine learning. Catboost [11] is an improved algorithm of GBDT algorithm. Compared with other Boosting algorithms, CatBoost has the advantage of efficiently processing categorical variables, effectively preventing overfitting, and having high model training accuracy.

4. Hybrid Model

In order to combine the advantages of traditional machine learning with deep learning, we propose the following hybrid model.

1) the PI Regression $X'$ (Rx) is obtained through Catboost algorithm regression.
2) Using the CGAN module to refine the Rx information, the purpose is to learn part of the correlation among the PI, Taking the Rx information as the input of the generator in the CGAN module
3) Using the discriminator to determine whether the input is noise data or not. In order to increase the stability of training. Using Loss (1) to constrain the generated results.

$$L_{mse} = \frac{1}{N} \sum (y_i - h_\theta (x_i))^2$$ (1)
At the same time, a classifier is added to optimize the classification performance of the generated data. Finally, the result of the generator is cascaded with the regression result, and then it is classified after training.

In the hybrid framework the form of the loss function is as (2):

\[
\text{Loss}(D, G) = L_{mse}(X'/G, Y) + L_{BCE}(X' \text{ Prob} / G, Y) + L_{BCE}(0 / 1, \text{ Noise} / G)
\]  

(2)

The hybrid framework as follows:

![Figure 1: The hybrid framework architecture](image)

\(X\): non-PI  
\(X'\): PI  
\(Y\): Label  
\(R\): Use catboost to regress \(X'\) one by one  
\(C1\): Use Catboost to classify PI and generate probability distribution about \(Y\)  
\(C2\): Catboost is used to classify the cascaded data  

5. Experimental comparison

In order to evaluate the performance of the proposed method, the experimental parameters are set as follows:

\(R\): iterations : 100000, learning rate : 0.1, loss function : 'RMSE', bootstrap type : 'MVS'

\(C1\): iterations : 100000, learning rate : 0.01, loss function : "MultiClass"

\(C2\): iterations : 50000, learning rate : 0.01, loss function : "MultiClass"

\(CGAN\):

epoch: 100, batch size: 8000, learning rate = 0.01

G: 6 layers fully connected network
D4, D2: 2 layers fully connected network

For the CGAN module, a generator(G) and a four classification model(D4) are trained by using neural networks. The accuracy and loss of the training process of the CGAN module have proved the advantages of this module in Fig.2 and Fig.3, which can be trained to verify the advantages of the algorithm.

Compared with training classifiers and generators separately, the CGAN module can significantly improve the performance of the model.

Furthermore, by comparing the effects of different merging methods of data on the C2 classifier, we have proved that the data generated by the CGAN module can effectively improve the final classification accuracy. We implement the model in Fig.1 and the result as Tab.1.

| EPOCH | 2000 | 5000 | 10000 | 20000 | 50000 |
|-------|------|------|-------|-------|-------|
|       | Train | Test | Train | Test | Train | Test | Train | Test | Train | Test |
| a     | 0.5757 | 0.5427 | 0.6161 | 0.5583 | 0.6496 | 0.5697 | 0.6919 | 0.5880 | 0.7489 | 0.5977 |
| b     | 0.9969 | 0.9813 | 0.9995 | 0.9843 | 0.9999 | 0.9837 | 1.0000 | 0.9837 | 1.0000 | 0.9840 |
| c     | 0.7481 | 0.6100 | 0.7771 | 0.6273 | 0.7985 | 0.6307 | 0.8098 | 0.6277 | 0.8127 | 0.6250 |
| a+c   | 0.7496 | 0.6117 | 0.7811 | 0.6253 | 0.8023 | 0.6290 | 0.8107 | 0.6247 | 0.8128 | 0.6250 |
| c+c   | 0.7478 | 0.6110 | 0.7766 | 0.6263 | 0.7981 | 0.6320 | 0.8094 | 0.6260 | 0.8127 | 0.6237 |
| a+a   | 0.5775 | 0.5450 | 0.6160 | 0.5557 | 0.6526 | 0.5717 | 0.6910 | 0.5887 | 0.7487 | 0.5977 |
| d     | 0.7387 | 0.6057 | 0.7687 | 0.6227 | 0.7938 | 0.6250 | 0.8063 | 0.6253 | 0.8126 | 0.6253 |
| c+d   | 0.7489 | 0.6107 | 0.7773 | 0.6267 | 0.7999 | 0.6310 | 0.8095 | 0.6267 | 0.8128 | 0.6240 |
| a+c+d | 0.7493 | 0.6140 | 0.7819 | 0.6297 | 0.8013 | 0.6323 | 0.8110 | 0.6280 | 0.8128 | 0.6257 |
| a+d   | 0.7384 | 0.6060 | 0.7746 | 0.6227 | 0.7988 | 0.6253 | 0.8102 | 0.6250 | 0.8128 | 0.6227 |

Tab.1 the comparison results
Note: a: Non-PI data  b: PI data  c: Regression PI data  d: CGAN data

The regressor is widely used in the hybrid model. It is different from other models, we introduce a pluggable CGAN module to refine the regression $X'$. We believe that the CGAN data (d) will perform well in the accuracy because the joint features of training data will be learned in the CGAN module. At the same time, we also believed that a reasonable combination of data can further improve the performance of our model.

The experimental results show that the network can get the best result (63.23%) when the number of iterations is 10000, and the accuracy of the model can be effectively improved by adding the generated data (d). Compared with the best result (63.07%) by using regression PI (c). The accuracy of the best result by using the regression PI (c) is increased by 0.16%, and compared with the best result without using PI (a) (59.77%) is improved by 3.46%. In order to prove the effectiveness of CGAN data after refactoring, we design a comparative experiment (c) + (c) and (a) + (a), which proves that invalid data cascade cannot significantly improve the performance of the model, and further proves that CGAN module can generate more efficient data.

6. Conclusion
In order to effectively use the product quality data in the process of workpiece manufacturing, this paper proposed a quality prediction model for the workpiece manufacturing process which differs from the traditional regression methods. The proposed model uses the CGAN module to learn the joint distribution and generate high-quality’s data. The effectiveness of the hybrid model is proved by experiments. The model combines the advantages of deep learning with machine learning, the experimental results show the advantage of the new hybrid model.

Acknowledgment
This work was supported in part by the National Nature Science Foundation of China under Grant 61773166, Grant 61772369, and Grant 61672240; in part by the Joint Funds of the National Science Foundation of China under Grant U18092006; in part by the Shanghai Municipal Science and Technology Committee of Shanghai Out-Standing Academic Leaders Plan under Grant 19XD1434000; and in part by the 2030 National Key AI Program of China under Grant 2018AAA0100500; and in part by Project of Shanghai International Cooperation of Science and Technology under Grant 19490712800.

References
[1] Zhiqiang G, Qingchao M, Yongming H, Qin W and Ouyang Z 2020 An Improved Extreme Learning Machine Based on Auto-Encoder for Production Predictive Modeling of Industrial Processes IEEE 8th Data Driven Control and Learning Systems pp 708-712
[2] Kirchen I, Vogel-Heuser B, Hildenbrand P, Schulte R, Vogel M, Lechner M and Merklein M 2016 Data-driven model development for quality prediction in forming technology IEEE 15th Int. Conf. on industrial informatics pp 775-780
[3] Minchala-Avila L I, Remoso-Avecillas M and Sánchez C 2016 A comparative study of black-box models for cement quality prediction using input-output measurements of a closed circuit grinding Annual IEEE Systems pp 1-7
[4] Mohammadi P and Wang Z J 2016 Machine learning for quality prediction in abrasion-resistant
material manufacturing process IEEE Canadian Conf on Electrical and Computer Engineering pp 1-4
[5] Purr S, Wendt A and Meinhardt J 2016 Data-driven inline optimization of the manufacturing process of car body parts IOP Conf Series: Materials Science and Engineering 159(1) pp002-012
[6] Ragab A, El-Koujok M and Poulin B 2018 Fault diagnosis in industrial chemical processes using interpretable patterns based on Logical Analysis of Data Expert Systems with Applications 95 pp368-383
[7] Han Q, Li H and Dong W 2017 On fault prediction based on industrial big data 36th Chinese Control Conf pp 10127-10131
[8] Cherukuri H, Perez-Bernabeu E and Selles M 2019 Machining chatter prediction using a data learning model J. Manufacturing and Materials Proc 3(2) pp45-50
[9] Jiang T, Zeng J and Zhou K 2019 Lifelong Disk Failure Prediction via GAN-Based Anomaly Detection IEEE 37th Int. Conf. on Computer Design pp 199-207
[10] Mirza M and Osindero S 2014 Conditional generative adversarial nets arXiv preprint p1411 p1784
[11] Prokhorenkova L, Gusev G, Vorobev A and Gulin A 2018 CatBoost: unbiased boosting with categorical features Advances in neural information Proc. systems pp6638-6648