ATLANTIS PRESS
Advances in Computer Science Research, volume 93
2nd International Conference on Mathematics, Modeling and Simulation Technologies and Applications (MMSTA 2019)

Forecast of Production Quantity of General-Purpose Parts Based on Customized Production

Honglin Wang[1], Xiaohong Li[2], Yongfeng Xi[1] and Tao Hong[1,*]
[1]University of Electronic Science and Technology of China, China
[2]Sichuan Deen Precision Co., Ltd., China
[*]Corresponding author

***Abstract*—With the development of the Internet, informatization has penetrated into all walks of life. The general-purpose parts manufacturing enterprise is no exception, and its production mode is also undergoing tremendous changes. The main performance is the development of customized production to the network. The production line of the enterprise has shifted from large-scale mass production to small-volume customized production. In this transformation process, a large amount of product quality related data is generated. Traditional quality data analysis methods have been difficult to unearth the full value of data.**

Therefore, this paper applies data mining technology to the quality-related data of general-purpose parts manufacturing enterprises. Through the prediction algorithm, it can accurately estimate the quantity of materials before production, which can reduce the inventory of enterprises or avoid secondary production. Due to the characteristics of the manufacturing process of the common parts, there are many reasons for the defects that may occur in the production, and even the same product, the total number of defects and the specific defects that are generated by different production batches are not the same. In the face of such a scenario, the usual prediction algorithm cannot be directly applied, because there is no feature vector as the algorithm input, so this paper proposes a comprehensive prediction algorithm based on regression-association-cluster to solve this problem. The algorithm only needs the product coding and the number of materials to be fed, so it can predict the number of scraps of the production task through historical data, and combine the production line experience to improve the accuracy of the prediction. Experiments show that the prediction effect of the algorithm is in line with expectations. The research work of this paper is supported by Sichuan Science & Technology Program under Grant No. 2018GZ0118.

Keywords—association rules; FP-growth; equality regression;clustering; K-means

I. Introduction

Nowadays, a variety of information systems are used within the general-purpose parts manufacturing enterprise, and a large amount of data is accumulated in the system database. Traditional data analysis methods are aimed at a small amount of data, and with the rapid growth of data, it can no longer be applied. In recent years, data mining technology has developed rapidly, and its application cases cover many fields such as business management, market analysis, financial analysis, medical diagnosis, etc., and its forecasting and analysis results are good. Data mining refers to the process of discovering the rules from a large number of disorganized data and then summarizing the knowledge. The knowledge gained can help companies make decisions.

This paper explores the application of data mining techniques to product quality related data for general-purpose parts manufacturing companies. General-purpose parts manufacturing companies are moving toward a customized production model of the network, transforming from a manufacturing-oriented company to a service-oriented enterprise. Under this model, the enterprise information system will accumulate a large amount of product quality related data, and the data mining technology can predict the number of inputs produced by the enterprise, and provide an auxiliary decision-making basis for the production plan of the enterprise, which is meaningful for small-volume customized production enterprises.

II. Characteristics of Quality Data of General-Purpose Parts Manufacturing Enterprises

The quality data of the general parts manufacturing company is closely related to its production process and industry characteristics. The following article gives a sample of the quality data of a general-purpose manufacturing company, as shown in Table 3-1. The sample data is convenient for this paper to introduce the characteristics of enterprise quality data, and the real data is more complicated than the sample data. The meanings of the main fields in Table 3-1 are as follows:

- Primary key encoding: Each primary key encoding code is a specific product.
- Planned number: The number of products expected to be produced. Number of materials to be picked: The amount of raw materials received, because the product will be scrapped during the production process, so the number of materials to be picked is more than the number of plans.
- Print order number: A print order number represents a production task and is the smallest unit of production.
- Discover the workshop: the workshop where the production is specific.
- Process No: The specific process that represents the occurrence of the scrap.

- Reason for the defect: The specific reason for the occurrence of this scrap.

TABLE I. SAMPLES OF MANUFACTURING DATA FOR GENERAL-PURPOSE PARTS MANUFACTURING COMPANIES

Line number	Main building code	Main part weight	Number of plans	Number of picks	Print order number	Discovery workshop	employee ID	Cause of defect	Number of failures
1	Cp01	5	500	600	1231	1	1	1	5
2	Cp01	5	500	600	1231	1	1	1	6
3	Cp01	5	500	600	1231	1	2	3	4
4	Cp01	5	500	600	1231	1	2	4	5
5	Cp01	5	1000	1300	1232	1	1	1	8
6	Cp01	5	1000	1300	1232	1	2	5	10
7	Cp01	5	1000	1300	1232	1	3	4	9
8	Cp02	3.2	300	350	1233	2	5	1	2
9	Cp02	3.2	300	350	1233	2	5	9	3
10	Cp02	3.2	300	350	1233	2	6	10	1
11	Cp02	3.2	5000	6000	1234	3	5	12	30
12	Cp02	3.2	5000	6000	1234	3	7	3	60

A specific record in Table 3-1, such as the first row record, represents the product Cp01 in the production task 1231, in the workshop 1, the process 1, the product was scrapped, the number of scrapped items was 5, and the reason for the scrap was the defect cause 1.

III. PREDICTION OF SCRAP AGE BASED ON REGRESSION-ASSOCIATION-CLUSTERING

Due to the industry characteristics of general-purpose parts manufacturing companies, there are many types of products. Different products, the specific processes and processes of their production are different. The manufacturing process is from scrap iron to finished product. There are many processes in the whole manufacturing process, and each process may be scrapped, and there are many external factors affecting the number of scrapped. These factors are not in the process of enterprise informationization. Collected, some can't even be quantified. However, these factors are precisely the decisive factors affecting the number of scraps. This brings a lot of uncertainty to the number of scraps in production tasks. Therefore, a comprehensive prediction algorithm based on regression-correlation-clustering is proposed to solve this problem.

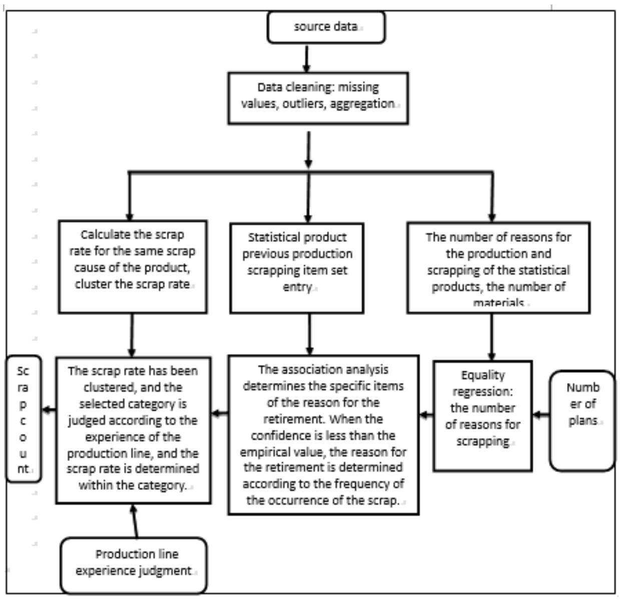

FIGURE I. DETAILED FLOW CHART OF COMPREHENSIVE PREDICTION ALGORITHM BASED ON REGRESSION-ASSOCIATION-CLUSTER.

The comprehensive prediction algorithm based on regression-association-cluster is divided into three stages, and the refinement process of each stage is shown in Figure I. After the source data is cleaned by the data, it supports the processing of each stage. The input of the whole algorithm consists of two parts, one is the planned number of plans to be produced, and the other is the judgment made by the production line based on the actual situation after the algorithm clusters the source data. In the first processing step, it is necessary to count the sequence of the number of production scrapping reasons of the product from the source data, and the number of the corresponding number of scrapping reasons, and then use the equation regression method to establish the model, and use the number of materials to predict The number of possible defects in this production. The number of defects is the basis of the second processing step, because the second step needs to determine the specific cause of the defect. Therefore, the second step needs to obtain the product production scrap item set from the source data. The second step process incorporates some of the idea of relevance mining, but if the correlation is weak, the confidence is less than an empirical value. Instead, select a single item set with a high frequency. This will determine the specific cause of the defect. The third step will sort out the historical sequence of product scrap rate of the product on each defect cause from the source data. Then cluster this sequence to find the best dividing point, rely on the production line to give an empirical judgment, and finally determine the number of product scraps.

Because the source data has about 500,000 pieces of data, it is decomposed into specific products, and the amount of data varies from dozens of data to several thousand pieces of data. In order to facilitate the following introduction, this paper selects a product with less data volume as sample data, and introduces the whole algorithm prediction process. The sample product has 265 records in the source data, and after data preprocessing, there are 57 valid data left.

A. Reasons for Predicting Product Defects

The first step in the algorithm is to predict the number of reasons why the product may be defective in this production task. Because from the historical data, it is possible to count the number of causes of defects in previous production tasks, and the corresponding number of inputs. Here, the number of defects in the i-th production is recorded as the number of corresponding inputs, so the first step is to establish a function of x and y so that the new input x can predict y well. The correspondence table of x and y calculated from the preprocessed data is given below, as shown in Table II.

TABLE II. STATISTICS ON THE CAUSES OF HISTORICAL DEFECTS

y	3	2	2	2	1	1
x	3573.3929	1204.75	1193.75	122.5	122.5	520
y	1	2	1	2	2	3
x	113	1506.75	1004.75	1002.75	1005.25	1512
y	1	1	1	1	6	2
x	3086.75	1005	1007.75	2513.25	2062.25	151.5
y	1	2	1	1	5	2
x	1006.25	1050.25	518	310.5	6267.5	1007.25
y	1	1	1	2	1	1
x	506.75	606	1036.5	1088	2007.75	1007.5

The algorithm uses regression techniques to find functions for x and y. There are many kinds of regression techniques, and linear regression is not very suitable here, because it can be seen from the data that the range of the x-axis is very large, the approximate range is [1,10000], and the variation range of the y-axis is small. Probably [1,6]. This is a good understanding, the number of plans can be very large, because there are large orders, and the cause of the defects is limited, generally no more than 10 defects in a product. So this should be a curve, that is to say a nonlinear fit.

The ridge regression also has problems here. The curve of the ridge regression fitting is given below, as shown in Figure 3-6. For ease of presentation, the x-axis data in Figure 3-6 is 100 times smaller than the real data. It can be seen that the ridge regression does fit a curve, but the curve is the trend of rising first and then falling. This is because historical data is limited compared to the x-axis range and does not completely cover the entire x-axis. In addition, there are occasionally high data in historical data, or abnormal data caused by manual entry errors, such as points in Table II (2062.25, 6). Such points have a large interference with ridge regression, resulting in curves. The trend is abnormal, and the general law cannot be well characterized.

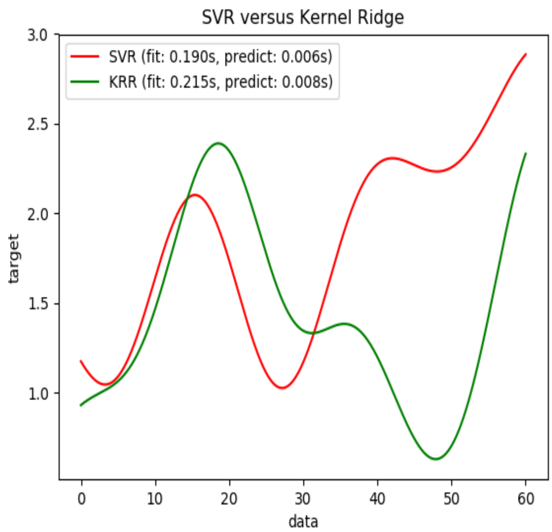

FIGURE II. NUCLEAR RIDGE REGRESSION CURVE

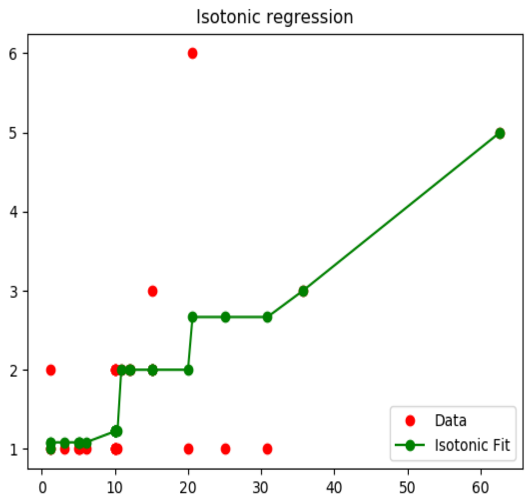

FIGURE III. EQUIVALENCE REGRESSION CURVE

Compared to the ridge regression, the equation regression does not have this problem. This paper first gives the curve of the equation regression fitting, as shown in Figure 3-7. Similarly, for ease of presentation, the x-axis data in Figure 3-7 is 100 times smaller than the real data. It can be seen from the figure that the outliers do not cause interference to the trend of the curve, and the curve shows a trend of rising in stages. The equality regression fits the non-descending function of the data, so this paper chooses the equality regression as the regression function. Suppose there is an order with a planned number of 2000, and 2100 is used as the input of the number of inputs. The equation is used to predict the number of defects, and the value of the output of the function is rounded up to obtain the number of defects. The number of defects is 3. 3 defects The reason will be the basis for the next step.

B. Reasons for Predicting Product Scrapping

From the previous step, it is expected that there will be three defects in this production. So now it's time to determine the specific code for the three causes of the defect. From the historical data, you can know the specific causes of defects in previous productions. This article refers to the collection of the causes of defects in each production task as an instance. In the step of determining the cause of the specific defect of the product, the idea of some association rules is borrowed, but slightly different. The defect cause data set of this example is first given below, as shown in Table III.

TABLE III. DEFECT CAUSE ITEMS

Serial number	Item set	Serial number	Item set
1	173	17	219
2	101	18	145、106、219
3	130、173	19	219
4	219、101	20	219
5	222	21	173
6	130	22	173
7	152、116、173	23	219、102
8	219	24	101
9	102	25	152、101
10	155	26	155、222
11	152、106	27	128
12	152、102、219、173、222	28	222
13	152、219	29	219
14	116、102	30	102
15	152、219、222	31	130、173、116、118、152、219
16	152	32	219、222

First, you need to scan the data set for the first time, count the single item set, and sort the single item set according to the frequency of occurrence. The sorting results are shown in Table 3-6. Here, the paper only gives some results with higher frequency.

TABLE IV. FREQUENCY OF DEFECTS

Cause of defect	occurrence frequency
219	13
152	8
173	7
222	6
102	5
101	4
130	3
116	3

In previous productions, the most frequent cause of defects occurs, and the possibility of recurrence in the next production is high. So the paper adds the most frequent single item set 219 to the result set of the second step. With the first cause of the defect, using the correlation to find the second cause of the defect, when the confidence of the correlation is greater than a certain empirical value, it can be considered that the two causes of the defect are related, and the probability of occurrence is high, so It is possible to determine other causes of defects. When the confidence level is lower than the empirical value, it indicates that the cause of the defect in the product is not strong. The paper selects the defect reason and puts it into the result set according to the frequency of occurrence of the single item set. This is to make the final selection of the cause of the defect more extensive and improve the accuracy of the prediction.

Finally, the reasons for the defects selected in this paper are: 219, 152, 173.

C. Forecasting Product Retirement

The final step of the algorithm is to cluster the historical data for each defect cause for the cause of the defect obtained in the previous step. The object of clustering in this paper is the scrap rate of previous production, which is scrapped in the cause of this defect.

There are many algorithms for clustering to target different application scenarios. In this paper, the scrap rate is clustered, which is one-dimensional data, which can be converted into two-dimensional data. This article is more concerned with the Euclidean distance between points, so choose the K-Means algorithm for clustering.

The following paper first gives the historical scrap rate for defect cause 219, as shown in Table V.

TABLE V. COMPREHENSIVE PREDICTION ALGORITHM 0~500 SOURCE DATA VOLUME TEST RESULT DEFECT CAUSE 219REQUENCY IN

Serial number	Scrap rate	Serial number	Scrap rate
1	1.23132%	8	0.26455%
2	0.996057%	9	0.259172%
3	3.26531%	10	0.387926%
4	3.53982%	11	0.795031%
5	0.265472%	12	0.0638213%
6	1.19671%	13	0.397022%
7	0.397911%		

Then use the K-Means algorithm to cluster the scrap rate of defect cause 219, as shown in Figure 3-8. In order to facilitate the display, the data of the x-axis is expanded by 100 times, the x-axis represents the scrap rate, and the y-axis has no practical significance.

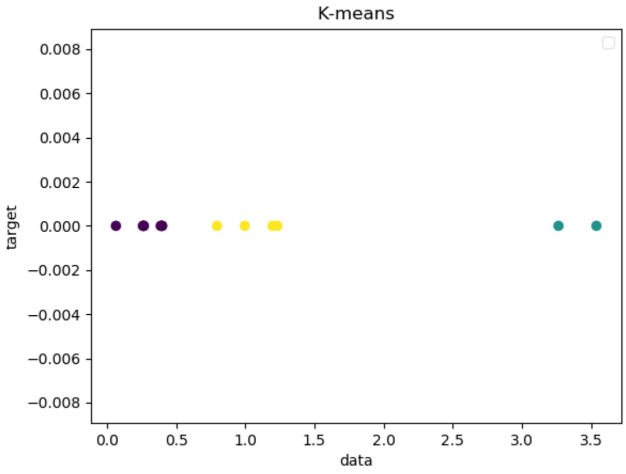

FIGURE IV. DEFECT CAUSE 219 CLUSTERING

As can be seen intuitively from the figure, clustering is very obvious. It can be seen that the cause of the defect is divided into three categories, and the scrap rate changes stepwise as the class changes. It is assumed here that the production line is in this production, and the rejection rate of the defect cause 219 is judged to be the first category, that is, the leftmost one of the x-axis, and then the record with the closest number of feeds to our current production is found from this class, corresponding to The scrap rate is the scrap rate predicted by the paper for the cause of defects 219. The scrap rate obtained here is 0.387926%, so the final forecast is 2,100 pieces of defects and 8 scraps on the cause of defects. Finally, a similar operation can be performed on the other two causes of the defect to output the final predicted result.

D. Analysis of experimental results

For the experiment of the scrap number prediction algorithm, this paper uses the simulated production environment to predict the scrapped quantity of a production task. The rules used in the paper experiment are as follows: the data of the last production task in the historical data of the product is removed as the training data of the algorithm. After the model is obtained, the number of inputs of the last production task and the product code input algorithm are calculated, and the experience judgment of the production line is simulated, and the predicted number of scrapped numbers of the current forecast is obtained, and the predicted scrapped number is compared with the actual number of scrapped occurrences.

The results of the experiment are shown in the table below.

TABLE VI. COMPREHENSIVE PREDICTION ALGORITHM 0~500 SOURCE DATA VOLUME TEST RESULT TATISTICS ON THE

product code	Number of feeds	Predictive value	actual value	Source data entry
BCP026371	890	89	75	210
BCP060532	2100	20	16	292
BCP000412	279	30	29	145
BCP002947	1131	167	81	357
BCP002959	873	34	21	305
BCP002976	900	38	22	205
BCP004379	1450	19	14	251
BCP006839	600	121	112	386
BCP004458	1272	54	55	365
BCP006025	266	28	29	421

TABLE VII. COMPREHENSIVE PREDICTION ALGORITHM 500~900 SOURCE DATA VOLUME TEST RESULTS OMPREHENSIVE PREDICTION

product code	Number of feeds	Predictive value	actual value	Source data entry
BCP004391	883	5	4	534
BCP006029	4453	185	172	759
BCP006013	30	6	5	796
BCP049772	180	33	30	624
BCP046028	500	54	50	791
CP005063	2040	52	28	645
BCP054229	5200	86	93	611
CP004935	1857	162	150	854
BCP041148	3521	343	266	802
BCP031256	1250	30	25	578

TABLE VIII. TEST RESULTS OF SOURCE DATA VOLUME ABOVE 900 FOR COMPREHENSIVE PREDICTION ALGORITHM

product code	Number of feeds	Predictive value	actual value	Source data entry
BCP006050	18922	1654	1300	939
BCP023407	850	50	41	1164
BCP004402	1479	111	130	917
BCP006048	6249	427	282	1817
BCP006917	2130	130	98	1100
BCP008490	130	9	9	1015
BCP008536	333	19	33	965
BCP016684	5911	734	615	1166
BCP005032	6200	358	287	1034
BCP024608	3216	267	225	1312

Among them, Table VI is the test result of the source data amount of 0~500. Samples with less than 100 source data are not sampled because this is rare. The paper has carried out sampling tests on the products of other data segments, and it can be found that the prediction effect in most cases is good.

Table VII shows the test results of the source data volume between 500 and 900. The sampling is relatively uniform. The test also showed that the predicted value deviated from the true value, and the prediction results of other tests performed well. Table VIII shows the test results with the source data volume greater than 900. The predicted values are close to the true values. From the overall experimental results, it can be found that the prediction effect of the algorithm is relatively stable, which is very meaningful for the practical application of the algorithm, and the enterprise cannot accept an algorithm that often fails to predict. In addition, the predicted value of the algorithm is close to the real value, and has reached the expectation. For a tens of thousands of orders, hundreds of pieces will eventually be produced, which is within the acceptance of the enterprise.

IV. SUMMARY

When first discusses the specificity of the quality data of the general parts manufacturing enterprise, which is the necessary preparation for mining the data. The prediction of the number of scraps, the paper uses a new idea, because of the characteristics of the manufacture of common parts, there are many factors influencing the forecast, and the enterprise has no historical quantitative data of relevant factors, and the existing data cannot be accurately predicted, so it is introduced. Production line experience to overcome this difficulty, the specific processing flow is described in detail in this article.

ACKNOWLEDGMENT

The research work of this paper was supported by Sichuan Science & Technology Program under Grant No. 2018GZ0118.

REFERENCES

[1] Dong Z, Yang D, Reindl T, et al. Satellite Image Analysis and a Hybrid ESSS/ANN Model to Forecast Solar Irradiance in the Tropics[J]. Energy Conversion and Management, 2014(79): 66-73.

[2] Chang B R, Tsai H F. Novel Hybrid Approach to Data-Packet-Flow Prediction for Improving Network Traffic Analysis[J]. Applied Soft Computing, 2009, 9(3): 1177-1183.

[3] Firat M, Gungor M. Generalized Regression Neural Networks and Feed Forward Neural Networks for Prediction of Scour Depth around Bridge Piers[J]. Advances in Engineering Software, 2009, 40(8): 731-737.

[4] Jordan MI. A Parallell Distributed Processing Approach[Z], 1986.

[5] Ilow J. Forecasting Network Traffic Using FARIMA Models with Heavy Tailed Innovations[C]IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2000(6): 3814-3817.

[6] Sohn S Y, Lim M. Hierarchical Forecasting Based on AR-GARCH Model in a Coherent Structure[J]. European Journal of Operational Research, 2007, 176(2): 1033-1040.

[7] Chen C, Hu J, Meng Q, et al. Short-Time Traffic Flow Prediction with ARIMA-GARCH Model[C]Intelligent Vehicles Symposium (IV), IEEE, 2011: 607-612.

[8] Reyes J, Morales-Esteban A, Martínez-Álvarez F. Neural Networks to Predict Earthquakes in Chile[J]. Applied Soft Computing, 2013, 13(2): 1314-1328.

[9] Singhal D, Swarup K S. Electricity Price Forecasting Using Artificial Neural Networks[J]. International Journal of Electrical Power & Energy Systems, 2011, 33(3): 550-555.

[10] Elman J L. Finding struture in time[J]. Cognitive Science, 1990, 14(90):179--211.

[11] Hochreiter S. Untersuchungen Zu Dynamischen Neuronalen Netzen[Z], 2004.

[12] Graves A. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8):1735-80.