An Intelligent Early Warning Method of Press-Assembly Quality Based on Outlier Data Detection and Linear Regression

XUE Shanliang[*], *LI Chen*

College of Computer Science and Technology/College of Artificial Intelligence, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, P. R. China

(Received 15 May 2020; revised 15 June 2020; accepted 20 July 2020)

Abstract: Focusing on controlling the press-assembly quality of high-precision servo mechanism, an intelligent early warning method based on outlier data detection and linear regression is proposed. Linear regression is used to deal with the relationship between assembly quality and press-assembly process, then the mathematical model of displacement-force in press-assembly process is established and a qualified press-assembly force range is defined for assembly quality control. To preprocess the raw dataset of displacement-force in the press-assembly process, an improved local outlier factor based on area density and P weight (LAOPW) is designed to eliminate the outliers which will result in inaccuracy of the mathematical model. A weighted distance based on information entropy is used to measure distance, and the reachable distance is replaced with P weight. Experiments show that the detection efficiency of the algorithm is improved by 5.6 ms compared with the traditional local outlier factor (LOF) algorithm, and the detection accuracy is improved by about 2% compared with the local outlier factor based on area density (LAOF) algorithm. The application of LAOPW algorithm and the linear regression model shows that it can effectively carry out intelligent early warning of press-assembly quality of high precision servo mechanism.

Key words: quality early warning; outlier data detection; linear regression; local outlier factor based on area density and P weight (LAOPW); information entropy; P weight

CLC number: TP301 **Document code:** A **Article ID:** 1005-1120(2020)04-0597-10

0 Introduction

High-precision servo mechanism is widely used in intelligent machinery, which requires high quality and reliability. However the structure of high-precision servo mechanism is very complex and the assembly process is extremely complicated, which results in difficulties in assembly quality control. In this paper, the technologies of outlier data mining and linear regression are applied to quality control for the press-assembly process of high-precision servo mechanism, in which the raw dataset of displacement-force can be collected. An outlier data detection method is designed to preprocess the raw data of displacement-force in the process, and linear regression is used to figure out the relationship between assembly quality and press-assembly process. A displacement-force mathematical model is established and a qualified press-assembly force range is defined to monitor the assembly quality. Then an intelligent early warning method of press-assembly quality is proposed.

Quality control is the process of organizing related activities in accordance with quality requirements. Many scholars have conducted relevant studies on the quality control problems in the manufacturing process. In the field of quality control on intelligent manufacturing, the technology of dynamic

[*]Corresponding author, E-mail address: xuesl@nuaa.edu.cn.

How to cite this article: XUE Shanliang, LI Chen. An intelligent early warning method of press-assembly quality based on outlier data detection and linear regression[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2020, 37 (4):597-606.

http://dx.doi.org/10.16356/j.1005-1120.2020.04.010

monitoring and quality early warning for production machining workshop based on Android mobile terminal has been studied by Yin et al[1]. In addition, the multi-point real-time intelligent neural network prediction model and algorithm were researched to achieve dynamic monitoring which finally realized early warning[1]. BP neural network was used to establish an early warning model for abnormal events in aircraft assembly workshop and classified early warning was realized in Ref.[2]. Wu et al.[3] raised a quality control method for the assembly process of complex products based on digital twin. The Markov method was applied to predict the quality data and provide early warning based on predicted value[3].

The studies above focus on the existing quality control methods and the outlier data of quality control were not considered, which will result in the inaccuracy of quality control. Meanwhile, local outlier factor (LOF) is an effective outlier detection algorithm. It can be used to judge the outlier of an object according to the LOF measurement. Therefore, the idea of improving the LOF algorithm as a preprocessing for quality warning is considered in this paper, and then a method of outlier data detection and quality early warning according to local outlier factor based on area density and P weight (LAOPW) is designed and normal distribution is proposed to find out a more reasonable quality data control range of high-precision servo mechanism.

Outliers are also named as anomalies, novelties, deviations and exceptions[4-6]. In general, outlier data mining is a process of finding an effective method to mine data objects that meet the definition of outliers. It is widely used in medical insurance detection[7], credit card fraud detection[8], abnormal weather detection[9], etc. The concept of LOF, in which relative density is used to measure the degree of outliers of data objects, was first proposed by Breunig et al[10]. Fast outlier detection based on local density score (FLDS) was put forward in Ref.[11], in which k-nearest neighbors were used. The k-means algorithm was designed to segment the dataset to find out singular points, by which detection efficiency was improved. Through this method, the time complexity was reduced from $O(n^2)$ to $O(n^{1.5})$, and the calculation time is about 20 times faster than LOF. Data field theory and the concept of an average potential difference were applied to improve the detection accuracy of outlier detection algorithms in Ref.[12]. Detection quality was improved by enhancing the density-based spatial clustering of application with noise (DBSCAN)[13] clustering algorithm and local outlier factor based on area density (LAOF) to determine the outlier degree in Ref.[14].

1 Preprocessing of Intelligent Press-Assembly Outlier Data Based on LAOPW

In this paper, linear regression is used to deal with the relationship between the press-assembly quality and process of high-precision servo mechanism, then a displacement-force mathematical model of the press-assembly process is established and a qualified press-assembly force range is defined for early warning of assembly quality control. However, the analysis of linear regression will be seriously affected by outliers. Therefore, before establishing a regression model, the outlier data detection method for preprocessing should be applied to eliminate outliers so that a more reasonable quality control range is defined.

1.1 Weighted distance based on information entropy

Information entropy[15-16] is used to measure the uncertainty of random variables. The greater the amount of information is, the smaller the uncertainty and the entropy are. Otherwise, the smaller the amount of information is, the greater the uncertainty and the entropy are. Therefore, the outlier degree of a certain data object can be evaluated by the entropy value. Let $s(x)$ be the set of random variable x, and $p(x)$ represents the probability, then the information entropy $H(x)$ will be defined as

$$H(x) = -\sum_{x\in s(x)} p(x)\ln p(x) \tag{1}$$

Let $A=\{A_1, A_2, \cdots, A_n\}$ be the attribute set of data object, $A_i\,(i=1, 2, \cdots, n)$ divides A into $\{A_i\}$ and $A-\{A_i\}$ which denoted as $P_1=\{A_i\}$, $P_2=\{A_1, A_2, \cdots, A_{i-1}, A_{i+1}, \cdots, A_n\}$ and $P=\{P_1, P_2\}$. Then the calculation formula of the increment of information entropy $\Delta(A_i)$[14] is shown in Eq.(2), that is

$$\Delta(A_i)=H(A)-H(P) \tag{2}$$

$$H(P)=\sum_{m=1,2}\left(\frac{P_m}{A}\left(H(P_m)\right)\right) \tag{3}$$

where $\Delta(A_i)$ represents the information entropy change of set A after removing A_i. The larger $\Delta(A_i)$ is, the more the uncertainty of the dataset is reduced.

In order to enhance the effect of outlier attributes in distance measurement, attribute-weighted distance is used. Given two data objects $p=\{p_k \mid k\in[1, n]\}$ and $q=\{q_k \mid k\in[1, n]\}$, where n is the number of attributes and k is the index of attributes, then the weighted distance between them is

$$d(p,q)=\left[\sum_{k=1}^{n}\Delta(A_k)\cdot d(p_k,q_k)\right] \tag{4}$$

1.2 Definition of LAOPW algorithm

In traditional LOF algorithm, there are some definitions as follows.

(1) K distance. The k distance $d_k(p)$ of object p refers to the distance between p and the object which is the kth nearest to it.

(2) K distance neighborhood. The k distance neighborhood of object p is a set of all objects whose distance between itself and p is less than or equal to $N_k(p)$. It can be expressed as

$$N_k(p)=\left\{q\in D \mid d(p,q)\leqslant d_k(p)\right\} \tag{5}$$

(3) Reachable distance. The reachable distance of p relative to q is defined as

$$\text{reach\_dist}_k(p,q)=\begin{cases} d_k(p) & p\in N_k(q)\\ d(p,q) & p\notin N_k(q)\end{cases} \tag{6}$$

The calculation formulas of local reachable density and LOF are

$$\text{lrd}_k(p)=\frac{|N_k(p)|}{\sum\limits_{q\in N_k(p)}\text{reach\_dist}_k(p,q)} \tag{7}$$

$$\text{LOF}_k(p)=\frac{\sum\limits_{q\in N_k(p)}\frac{\text{lrd}_k(q)}{\text{lrd}_k(p)}}{|N_k(p)|} \tag{8}$$

It can be noted that the sparseness of different data objects is not considered in LOF algorithm. Therefore, P weight of the data object is used as reachable distance in the LAOPW algorithm, and set as the area radius to obtain the regional area instead of the distance sum. The maintenance of algorithm running efficiency and the improvement of detection effect are gained as a result of the redefinition of local density and local outliers.

(4) P weight. The P weight $W_k(p)$ of object p equals to the reachable distance $k(p, q)$, which is the sum of the distances between p and its k neighborhoods. It is defined as

$$W_k(p)=k(p,q)=\sum_{q\in N_k(p)}d(p,q) \tag{9}$$

(5) Local density. For a circle, let the data object p be the center and $k(p, q)$ be the radius, then the number of data points per unit area is defined as the local density of p, which is expressed as

$$\text{LAO}_k(p)=\frac{|N_k(p)|}{\pi\left(k(p,q)\right)^2} \tag{10}$$

(6) Local outlier factor. The local outlier factor $\text{LAOPW}_k(p)$ of p is defined as

$$\text{LAOPW}_k(p)=\frac{\sum\limits_{q\in N_k(p)}\text{LAO}(q)}{\text{LAO}(p)|N_k(p)|} \tag{11}$$

For a certain data object p, the smaller its $\text{LAO}_k(p)$ is, the greater the $\text{LAOPW}_k(p)$ is, and the higher the outlier of p is, the more likely it is the outlier.

1.3 Flow of LAOPW algorithm

The outlier detection algorithm based on LAOPW is described in Algorithm 1 and its flow chart is presented in Fig.1.

Algorithm 1 LAOPW

Input: raw dataset D, k

Output: outliers of data objects

```
calculate $\Delta(A_i)$;
for each $x_i \in D$
    calculate $d_k(x_i)$ and $N_k(x_i)$ ;
end for
```

```
for each $x_i \in D$
    calculate $P$ weight($x_i$) ;
    calculate LAO$_k$($x_i$) ;
    calculate LAOPW$_k$($x_i$);
end for
descending output LAOPW$_k$($x_i$).
```

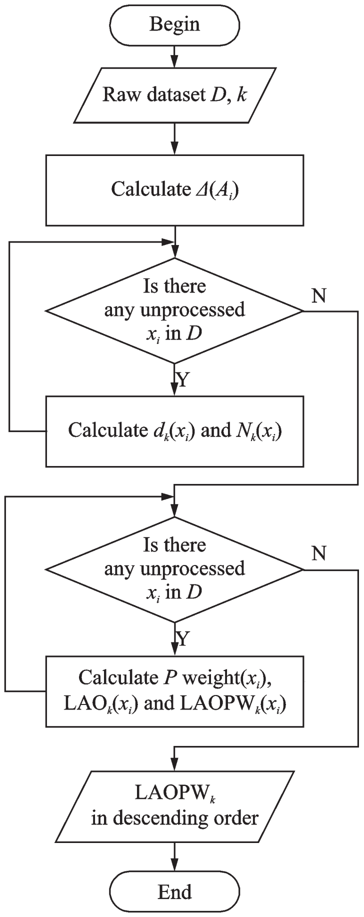

Begin
Raw dataset D, k
Calculate $\Delta(A_i)$
Is there any unprocessed x_i in D
N
Y
Calculate $d_k(x_i)$ and $N_k(x_i)$
Is there any unprocessed x_i in D
N
Y
Calculate P weight(x_i), LAO$_k$(x_i) and LAOPW$_k$(x_i)
LAOPW$_k$ in descending order
End

Fig.1 Flow chart of LAOPW algorithm

2 Intelligent Early Warning of Press-Assembly Quality Based on Normal Distribution

The detection method of outliers based on the normal distribution is a method based on statistics. Assuming that the given dataset is accordant with the normal distribution, and the data objects inconsistent with the model are identified as outlier data. If an attribute of a normal object is in accordance with the normal distribution $N(\mu, \sigma^2)$ (where μ and σ are the mean and standard deviation, respectively), it can be converted to the standard normal distribution $N(0, 1)$ by transforming $z=(x-\mu)/\sigma$, where μ and σ are unknown and can be estimated by the sample mean and standard deviation[17].

It can be seen from the law of large numbers that the normal distribution can be used to approximate other distributions when there are many samples. As shown in Fig.2, this theory can be applied to quality control. The middle solid black line μ is the predicted value of the observed value. $\mu \pm 2\sigma$ corresponds to the upper and lower warning lines, and $\mu \pm 3\sigma$ stands for the upper and lower control lines. If the distance between a sample and its mean μ exceeds 3σ, this value is identified as an outlier[17].

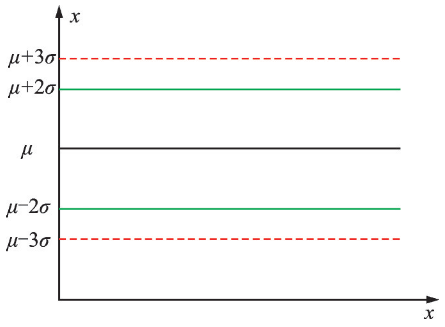


Fig.2 Schematic diagram of quality control

For a sample x, if there exists $\mu-2\sigma<x<\mu+2\sigma$, it means that the measurement process is under control and the production process is effective; If x meets the condition $(\mu+2\sigma \leqslant x<\mu+3\sigma) \| (\mu-3\sigma \leqslant x<\mu-2\sigma)$, it indicates that the quality is starting to deteriorate and tending to be “out of control”, so a necessary inspection should be carried out. If $(x \geqslant \mu+3\sigma) \| (x \leqslant \mu-3\sigma)$, it stands for “out of control” of the production process, the sample x is invalid or the product assembled is scrapped, therefore it should be checked and corrected immediately[17]. In this way, quality early warning for the intelligent press-assembly process can be realized.

3 Experiment and Application

To verify the outlier detection performance of LAOPW algorithm which is used to preprocess press-assembly data, two UCI datasets are used to compare and analyze several algorithms from multiple perspectives, they are the LOF algorithm, the

LAOF algorithm proposed in Ref.[14] and the LAOPW algorithm. All algorithms are implemented in Matlab with the experimental environment of Win10, and the processor is Intel(R) Core(TM) i5-8400 @ 2.80 GHz 2.81GHz.

3.1 Evaluation indexes

In the problem of outlier data detection, high detection effect is our pursuit and the detection performance of outlier data mining methods can be described by the confusion matrix[18] shown in Table 1.

Table 1 Confusion matrix

Confusion matrix		Predicted class	
		Outlier (O)	Normal (N)
Actual class	Outlier (O)	Outlier with true prediction (TN)	Outlier with wrong prediction (FP)
	Normal (N)	Normal with wrong prediction (FN)	Normal with true prediction (TP)

According to the relevant parameters of the confusion matrix, several indexes for evaluating the performance of outlier data mining algorithms are introduced.

(1) Accuracy. It represents the proportion of all samples that are correctly predicted, and stands for the overall prediction accuracy of the dataset. The larger the value is, the better it is. It can be expressed as

$$\text{Accuracy}=\frac{\text{TN}+\text{TP}}{\text{TN}+\text{FP}+\text{FN}+\text{TP}} \tag{12}$$

(2) Precision. It can be understood as how many of the data points are correctly predicted among those prediction results of normal categories. It is defined as

$$\text{Precision}=\frac{\text{TP}}{\text{FP}+\text{TP}} \tag{13}$$

(3) Recall. It is the rate of points that correctly predicted in all normal points. It can be written as

$$\text{Recall}=\frac{\text{TP}}{\text{FN}+\text{TP}} \tag{14}$$

(4) F-score. The formula of F-score is shown as

$$F\text{-score}=2\times\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{15}$$

Among the evaluation indexes above, Accuracy is used to measure the ability to make correct choices. Precision and Recall reflect the performance of outlier detection algorithms. And F-score is a comprehensive evaluation index of the two.

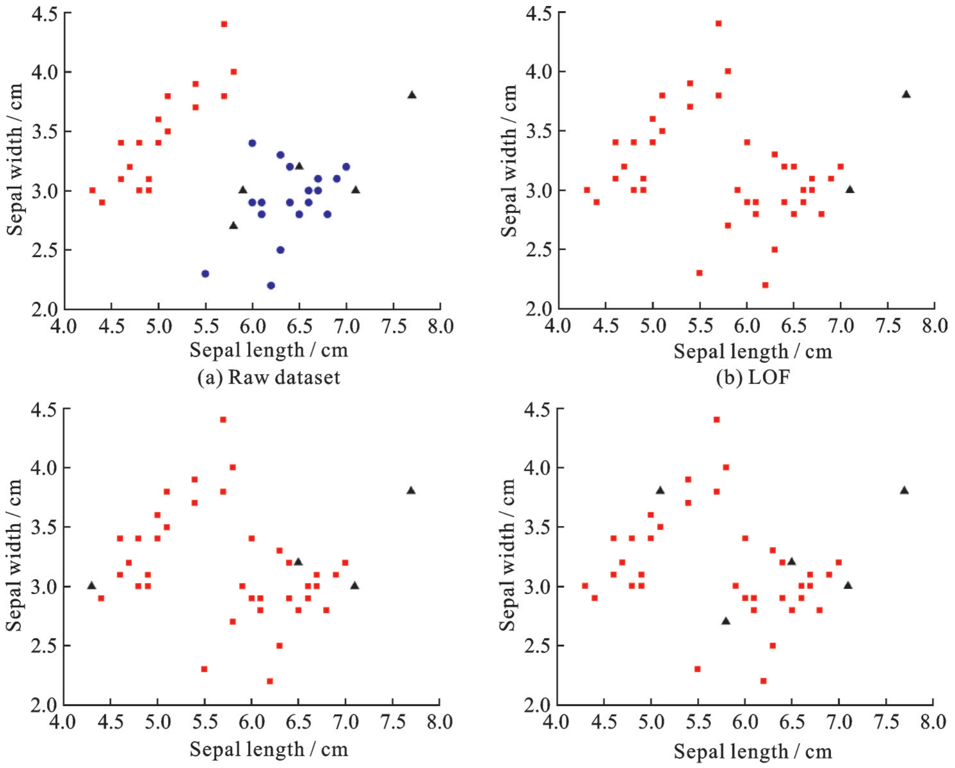
3.2 Experiment 1 (Iris dataset)

Iris dataset, which has 150 pieces of data and four attributes, is used for experiments of outlier data detection. The data objects are divided into three categories, including Setosa, Versicolour and Virginica. Twenty sample points belonging to Setosa and Versicolour are taken out as clusters and five Virginica sample points are selected as outlier data. Detection results of various algorithms are shown in Fig.3. The five black triangles in Fig.3(a) are outliers, and Figs.3(b)—(d) represent different detection results of comparison algorithms.

Based on Table 1, the confusion matrix of Iris for outlier data detection is listed in Table 2 and the histogram shown in Fig.4 is formed. It can be clearly seen from Table 2 and Fig.4 that the Accuracy, Precision and F-value of the algorithm for outlier detection in this paper are higher than those of LOF and LOAF algorithms. The Accuracy of LOF, which is 0.933, is the same as that of LOAF. Among them, the Recall values of the LAOF algorithm and the LAOPW algorithm are 0.975, which are slightly lower than that of the LOF. According to these indexes, the detection performance of the proposed algorithm is the best among these algorithms.

3.3 Experiment 2 (Aggregation dataset)

In order to compensate for the defect that the experimental data number in Section 3.2 is too small to fully demonstrate the effectiveness of the algorithm, Aggregation dataset is selected for comparative experiments. The data objects in this dataset are divided into seven categories, which are composed of 788 samples with a total of 2-D attributes. To carry out the experiment, 600 pieces of data



(a) Raw dataset

(b) LOF

(c) LAOF

(d) LAOPW

Fig.3 Experiment comparison based on Iris

Table 2 Confusion matrix for outlier data detection based on Iris

Confusion matrix	LOF		LAOF		LAOPW	
	O	N	O	N	O	N
O	2	3	3	2	4	1
N	0	40	1	39	1	39
Accuracy	0.933		0.933		0.956	
Precision	0.930		0.951		0.975	
Recall	1.000		0.975		0.975	
F-score	0.964		0.963		0.975	

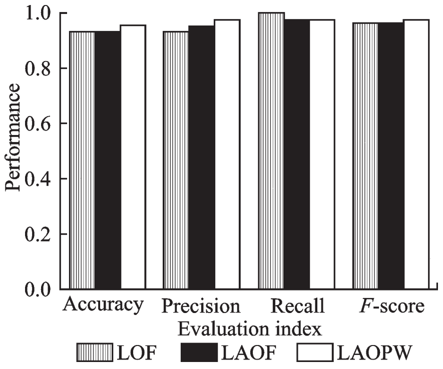
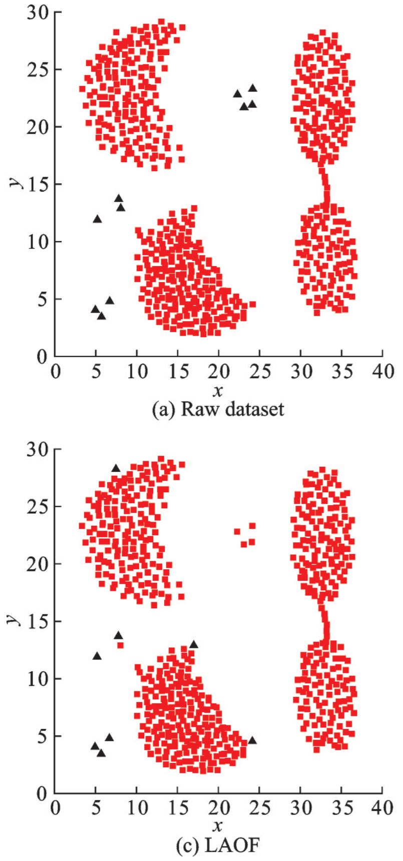
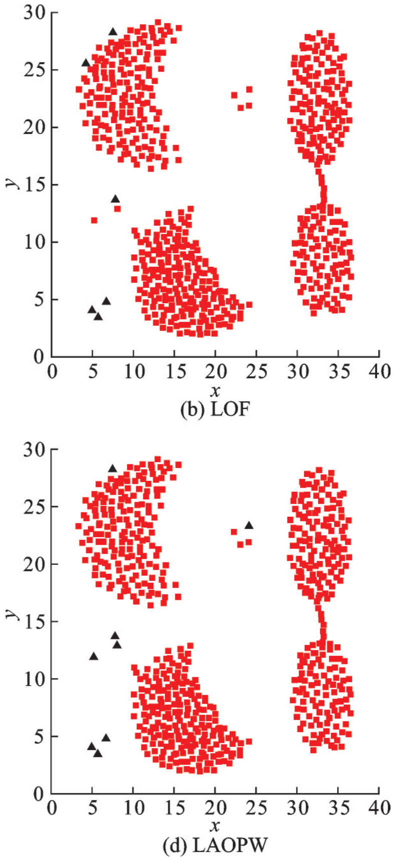


Fig.4 Evaluation index of different outlier data detection algorithms based on Iris

from four categories are extracted as cluster data, and 10 data points are selected from other three categories as outliers. The experimental results are shown in Fig.5. The detection performance of the algorithm is listed in Table 3 and Fig.6.

In Table 3, the number of outliers detected by the LOF algorithm is the smallest, only 6 and 2 of them are detected by mistake. By LAOF or the algorithm proposed, eight outliers can be detected, but the false detection rate of the latter is lower. The Accuracy of LAOPW is 0.993, which is higher than that 0.987 of LOF algorithm and LAOF algorithm. From Accuracy, Precision, Recall and *F*-score in Table 3 and Fig.6, the comprehensive detection performance of LAOPW is the best. This shows that the detection effect of the proposed algorithm is better than that of either the two algorithms.

In Aggregation dataset, 100, 200, 300, 400, and 500 data points are taken respectively to calculate the running time of the three methods. The results are shown in Fig.7. Among them, the LAOF algorithm has the highest operating efficiency, which is followed by the LAOPW algorithm, and the LOF algorithm is the lowest. The running time of the proposed LAOPW algorithm is slightly longer than that of the LAOF algorithm because the *P* weight of each object has to be calculated.

(a) Raw dataset

(b) LOF

(c) LAOF

(d) LAOPW

Fig.5 Experiment comparison based on Aggregation

Table 3 Confusion matrix for outlier data detection based on Aggregation

Confusion matrix	LOF O	LOF N	LAOF O	LAOF N	LAOPW O	LAOPW N
O	4	6	5	5	7	3
N	2	598	3	597	1	599
Accuracy	0.987		0.987		0.993	
Precision	0.990		0.992		0.995	
Recall	0.997		0.995		0.998	
F-score	0.993		0.993		0.996	

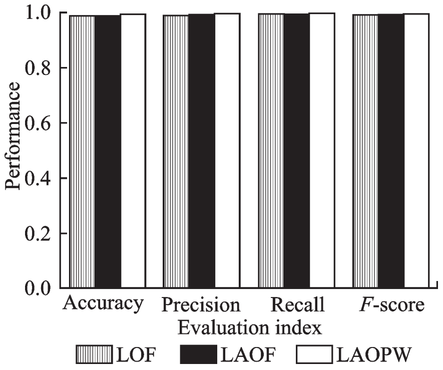


Fig.6 Evaluation index of different outlier data detection algorithms based on Aggregation

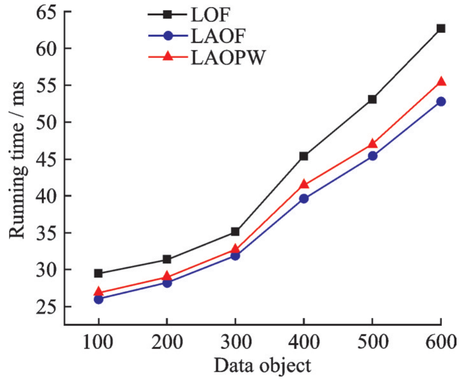


Fig.7 Comparison of running time among three different methods based on Aggregations

3.4 Application of intelligent early warning method of press-assembly quality

For the application of intelligent early warning method of press-assembly quality, the detection accuracy is more critical than the detection efficiency. The displacement-force raw data are collected from high-precision servo mechanism, as shown in Table 4.

The size of the displacement-force dataset is

Table 4 Displacement-force dataset of intelligent press-assembly

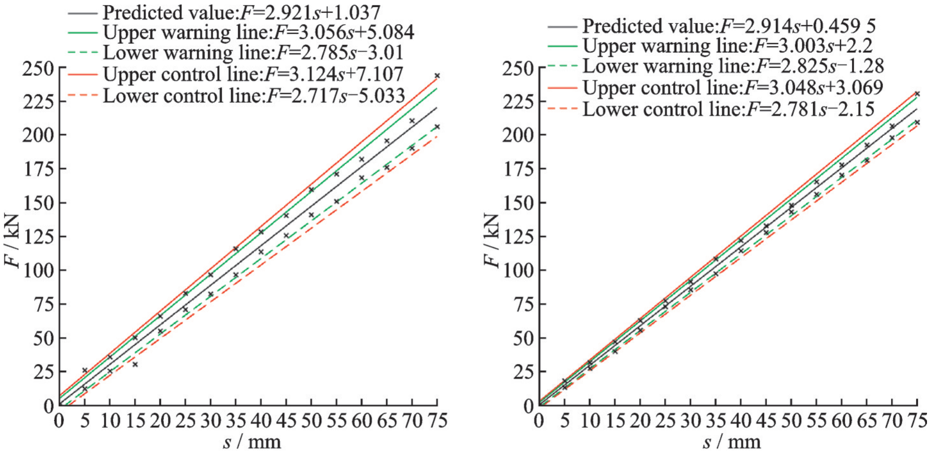
Displacement	Force F/kN			
s /mm	$F1$	$F2$	⋯	$F200$
0	0.000	0.000	⋯	0.000
5	14.352	17.456	⋯	15.102
10	28.324	28.705	⋯	27.438
15	44.055	42.577	⋯	42.184
⋮	⋮	⋮	⋮	⋮
70	206.257	203.112	⋯	204.664
75	233.275	225.095	⋯	222.827

200, and the number of attributes is 16. To eliminate the outlier data before linear regression model of displacement-force in press-assembly process is established and a qualified press-assembly force range is defined, the local outlier data detection algorithm LAOPW designed in this paper is applied to preprocess the raw dataset and eight outliers are detected, which are consistent with actual quality inspection results. In the new dataset without outliers, for each displacement value s, a univariate outlier detection method based on the normal distribution is used. Relevant statistical data is shown in Table 5.

The dependent variables μ, $\mu+2\sigma$, $\mu-2\sigma$, $\mu+3\sigma$, and $\mu-3\sigma$ have an approximately linear relationship with the independent variable s, so linear regression models can be established according to Table 5. Fig.8(a) represents the quality control chart from the raw displacement-force dataset, and Fig.8 (b) stands for the quality control chart by removing outliers with LAOPW algorithm. Since the collected displacement-force data points are too many, the symbol "×" is just used to identify the maximum and minimum forces under different displacements. Compared Fig.8(a) with Fig.8(b), it is obvious that the latter has a smaller quality control range and covers all data points. So it can be concluded that for a displacement-force dataset, a more accurate quality control range can be defined after removing outliers by the LAOPW algorithm, which can provide more reasonable control for the high-precision servo press-assembly process.

Table 5 Statistical data on displacement-force of intelligent press-assembly

s	Mean μ	Standard deviation σ	$\mu+2\sigma$	$\mu-2\sigma$	$\mu+3\sigma$	$\mu-3\sigma$
0.00	0.000	0.000 000	0.000 00	0.000 00	0.000 00	0.000 00
5.00	15.546	1.672 760	18.891 40	12.200 35	20.564 16	10.527 59
10.00	29.467	1.790 246	33.047 87	25.886 88	34.838 11	24.096 64
15.00	43.803	2.218 655	48.240 56	39.365 94	50.459 22	37.147 28
⋮	⋮	⋮	⋮	⋮	⋮	⋮
70.00	203.580	2.964 589	209.509 40	197.651 10	212.474 00	194.686 50
75.00	223.251	7.607 030	238.464 90	208.036 80	246.072 00	200.429 80



(a) Result based on raw dataset (b) Result after using LAOPW

Fig.8 Quality control chart of intelligent press-assembly process

Quality early warning can be realized by applying this control chart model to press-assembly process of high-precision servo mechanism. In such a process, $s=20$ mm and the corresponding force is F. If F is in the range between the upper warning line and lower warning line corresponding to s, it indicates that the press-assembly process works well. If it is within the two areas of the upper warning line and the upper control line, the lower warning line and the lower control line, it reveals that there is a problem with the press-assembly quality, and a necessary inspection and corresponding measures should be carried out. Once F is beyond the range between the upper control line and the lower control line, it means that the press-assembly process is abnormal, which may results in scraps. In the way above, the quality early warning for the high-precision servo mechanism press-assembly process can be realized.

4 Conclusions

An intelligent early warning method of press-assembly quality based on outlier data detection and linear regression is presented in this paper. Firstly, an improved outlier data detection algorithm LAOPW is designed for the preprocessing of press-assembly data. The experiments indicate that the proposed LAOPW algorithm has better comprehensive detection performance than LOF and LAOF algorithms. Then, the algorithm is used to preprocess the displacement-force data in the press-assembly process and the data objects with larger outlier factors are eliminated. Finally, the outlier detection method based on the normal distribution is applied to define the quality control range of the process, which is used as standard value of early warning for press-assembly quality. It can be used to monitor press-assembly process by collecting the force corresponding to different displacements. In this way, the problems of assembly quality can be found out in time and early warning can be given, then intelligent quality control will be further realized.

References

[1] YIN Chao, WANG Mingyuan, LI Xiaobin, et al. Real-time control support system of workshop production quality information based on mobile terminals[J]. Computer Integrated Manufacturing System, 2015, 21(1): 169-179. (in Chinese)

[2] SHAN Siqing, MAO Zhonghui, XIN Tenglong. Prediction technology of abnormal events in aircraft assembly workshop based on BP neural network[J]. Aeronautical Manufacturing Technology, 2014, 452(8): 42-47. (in Chinese)

[3] WU Ying, YAO Liya, XIONG Hui, et al. Quality control method of complex product assembly process based on digital twin technology[J]. Computer Integrated Manufacturing System, 2019, 25(6): 1568-1575. (in Chinese)

[4] HAN Jiawei, PEI J, KAMBER M. Data mining: Concepts and techniques[M]. [S.l.]: Elsevier, 2011: 543-583.

[5] CHANDOLA V, BANERJEE A, KUMAR V. Anomaly detection: A survey[J]. ACM Computing Surveys (CSUR), 2009, 41(3): 1-58.

[6] HODGE V, AUSTIN J. A survey of outlier detection methodologies[J]. Artificial Intelligence Review, 2004, 22(2): 85-126.

[7] RAWTE V, ANURADHA G. Fraud detection in health insurance using data mining techniques[C]//Proceedings of 2015 International Conference on Communication, Information & Computing Technology (ICCICT). [S.l.]:IEEE, 2015: 1-5.

[8] MALINI N, PUSHPA M. Analysis on credit card fraud identification techniques based on KNN and outlier detection[C]//Proceedings of 2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB). [S.l.]:IEEE, 2017: 255-258.

[9] OTSUKA T, TORII Y, ITO T. Anomaly detection algorithm for localized abnormal weather using low-cost wireless sensor nodes[C]//Proceedings of 2014 IEEE 7th International Conference on Service-Oriented Computing and Applications. [S.l.]:IEEE, 2014: 304-308.

[10] BREUNIG M M, KRIEGEL H P, NG R T, et al. LOF: Identifying density-based local outliers[C]//Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. [S. l.]: ACM, 2000: 93-104.

[11] THANG V V, PANTIUKHIN D V, NAZAROV A N. FLDS: Fast outlier detection based on local density score[C]//Proceedings of 2016 International Conference on Engineering and Telecommunication. [S.

l.]: IEEE, 2016: 137-141.

[12] MENG Haidong, SUN Xinjun, SONG Yuchen. Improved LOF algorithm based on data field[J]. Computer Engineering and Applications, 2019, 55(3): 154-158. (in Chinese)

[13] WANG W T, WU Y L, TANG C Y, et al. Adaptive density-based spatial clustering of applications with noise (DBSCAN) according to data[C]//Proceedings of 2015 International Conference on Machine Learning and Cybernetics (ICMLC). [S.l.]: IEEE, 2015, 1: 445-451.

[14] SHI Hongyan, MA Xiaojuan. Two-stage outlier detection method based on DBSCAN clustering and LAOF of hybrid data[J]. Journal of Chinese Computer Systems, 2018, 1: 74-77. (in Chinese)

[15] YANG Maolin. Research on outlier detection algorithm[D]. Wuhan: Huazhong University of Science and Technology, 2012. (in Chinese)

[16] ZHANG He, CAI Jianghui, ZHANG Jifu, et al. An outlier mining algorithm based on information entropy[J]. CAAI Transactions on Intelligent Systems, 2010, 5(2): 150-155. (in Chinese)

[17] XU Guogen, JIA Ying. Practical big data-detailed explanation and practice of MATLAB data mining[M]. Beijing: Tsinghua University Press, 2017: 480-487. (in Chinese)

[18] ZHU Lu. Improvement of DBSCAN algorithm based on adaptive estimation of EPS parameters and its application in outlier detection[D]. Kunming: Yunnan University, 2019. (in Chinese)

Acknowledgement The authors would like to acknowledge the following people for their assistance: WANG Qiquan and CHEN Guangxin from Nanjing Chenguang Group, and XIAO Xue, XU Meijiao, WU Liuyan, PENG Zhenfeng and LIU Xingxiong from College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics.

Author Dr. **XUE Shanliang** received the B.S. degree in mechanical engineering, the M.S. degree in mechanical manufacture and automation and the Ph.D. degree in mechanical and electrical engineering from Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 1992, 1999 and 2002, respectively. From 2015 to 2016, he was a visiting scholar in College of Electrical and Computer Engineering, Oklahoma State University, USA. From 2002 to present, he has been in the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, where he is currently an associate professor. His research is focused on intelligent manufacturing and computer application, which includes manufacturing network, data twins and Cyber-Physical System.

Author contributions Dr. **XUE Shanliang** designed the study, conducted the analysis, interpreted the results and wrote Section 0, Section 4 and abstract of the manuscript. Miss. **LI Chen** compiled the models, conducted data analysis and experiments, wrote Sections 1—3 of the manuscript. All authors commented on the manuscript draft and approved the submission.

Competing interests The authors declare no competing interests.

(Production Editor: SUN Jing)

一种基于离群数据检测和线性回归的压装质量智能预警方法

薛善良,李 晨

(南京航空航天大学计算机科学与技术学院/人工智能学院,南京 211106,中国)

摘要:针对高精度伺服机构压装质量控制难度大的问题,提出了一种基于离群数据检测和线性回归的智能质量预警方法。采用线性回归分析装配质量与压装过程之间的关系,建立了压装的“位移-力”数学模型,并定义了合格的压装力范围对装配质量进行控制。为了对压装过程中的“位移-力”原始数据集进行预处理,本文设计了一种改进的基于区域密度和P权值的局部离群因子(Local outlier factor based on area density and P weight, LAOPW)检测算法,以剔除导致线性回归数学模型不准确的离群值。该算法引入了基于信息熵的加权距离进行距离度量,并用P权值代替可达距离。实验结果表明,该算法在检测效率上比传统的局部离群因子(Local outlier factor, LOF)算法提高了5.6 ms,而检测准确率比基于区域密度的局部离群因子(Local outlier factor based on area density, LAOF)算法改善了2%左右。将本文提出的LAOPW算法和线性回归模型应用于高精度伺服机构压装质量控制,能够有效进行压装质量智能预警。

关键词:质量预警;离群数据检测;线性回归;基于区域密度和P权值的局部离群因子;信息熵;P权值