The Method of Anomaly Location Data Recognition Based on Improved YOLO Algorithm

Chen-can Wang[1(✉)], Yan Ge[2], and Yang Li[1]

[1] Information Engineering University, Zhengzhou 450000, China
[2] Zhengzhou Campus of Armyartillery Air Defense College, Zhengzhou 450000, China

Abstract. The existing anomaly location data recognition methods usually have poor accuracy due to the rough contour curve, so the anomaly location data recognition method is studied based on the improved YOLO algorithm. The improved YOLO algorithm is designed to judge the input and output residual error comparison in the normalization process. Based on the algorithm, the abnormal data location technology is studied, and the contour curve with low noise factor is obtained. Based on the improved YOLO algorithm, the abnormal location data recognition method is designed, and the accuracy of the method is optimized. The experimental results show that in the calculation of the first type error rate and the second type error rate, the slope of the method is gentle, and the value is small. It can be seen that the method will not produce large numerical changes under the changes of mathematical expectation and regression parameters, and can more accurately realize the anomaly location data recognition.

Keywords: Data exception · Improved Yolo algorithm · Exception recognition · Residual error

1 Introduction

For centralized data, if some data are very different from other data in structure, then they belong to abnormal data. Due to some errors occurring in the process of data collection and data transmission, abnormal data has little value, but for some special abnormal data, it has value that can be used, and some useful rules and knowledge can be obtained through their analysis. Therefore, how to remove these abnormal data has become an expert to solve the key problems [1]. Abnormal data greatly reduces the quality of data. If the quality of data is not guaranteed, there will be deviation in the follow-up data mining, and the analysis results such as clustering and regression can not be obtained. Therefore, a correct analysis report can not be provided to the decision-makers, which leads to deviation in prediction and decision-making, and ultimately leads to the loss of significance of the whole research. Therefore, more and more scholars are paying attention to this field, trying to find a systematic, complete and more adaptive anomaly data detection scheme.

S. Liu and X. Ma (Eds.): ADHIP 2021, LNICST 416, pp. 56–66, 2022.
https://doi.org/10.1007/978-3-030-94551-0_5

Reference [2] uses statistical concepts to construct a practical method, combined with simple formulas and visual representation of data to simplify the evaluation process. In reference [3], the structure of data existence is defined in the way of "object and attribute". Then the concept of similarity between theoretical distribution and attribute data distribution is proposed and analyzed, and the similarity algorithm of attribute data distribution is established. The advantage of using statistical methods to detect abnormal data in reference [4] is that it has enough statistical theory support, so the detected results are more reliable.

In this paper, through the above literature, based on the improved YOLO algorithm for anomaly location data recognition method is studied. Comparison of input and output residues during judging standardization. Based on this algorithm, the anomalous data localization technique is studied, and the profile curve of the low noise coefficient is obtained.

2 Research on Anomaly Location Data Recognition Method Based on Improved YOLO Algorithm

2.1 Design of Improved YOLO Algorithm

The YOLO algorithm, namely you only look once, means that you can identify the target you need at one glance. It is an end-to-end target detection algorithm, which has the huge advantage of fast detection speed. The main advantage of polo algorithm is that the monitoring effect is faster, and it can process video in real time in less than 25 ms. Compared with the traditional data location algorithm, this algorithm does not generate prediction box. In the case of convolution operation for each prediction box, a single regression method is designed for target detection, and the prediction box and confidence are obtained by deep convolution network processing directly from the input image. Instead of running on common frameworks, the algorithm uses Darknet framework, which uses C language and supports CPU and GPU. For target detection, Darknet can choose between two opencv and CUDA dependencies. Although Darknet is not as powerful as tensorflow and cafe, it is easy to install. Because it is written in C language, it is easy to check and debug the code. It can run without dependencies. It has Python interface, which is convenient to call the code and enhances the expansion ability of the framework. In addition, it can be well installed on personal computers, and has strong portability [5]. In addition, alexnet and vgg16 can also run on Darknet. Although there is no special team for maintenance, with these advantages, although they are relatively small, there is a lot of room for development. However, under the premise of extremely fast computing speed, the accuracy of this algorithm can not reach the ideal state, so this paper improves the algorithm, and designs the improved algorithm under different scales.

The improved YOLO algorithm mainly makes the following improvements: firstly, target detection of different scales, with 13×13, 26×26 and 52×52 as three different scale prediction selection values. In addition, the network structure is improved to some extent, using RESNET design method, adding residual blocks to reduce the possibility of gradient disappearance. The minimum scale feature map is responsible for the

detection of large objects, the medium scale feature map is responsible for the detection of medium objects, and the maximum feature map is responsible for the detection of small objects. The introduction of residual block reduces the parameter redundancy of convolution, reduces the situation of gradient disappearance and gradient explosion, and optimizes the overall network structure. For improved algorithms at different scales, the data input/output process is shown at Fig. 1.

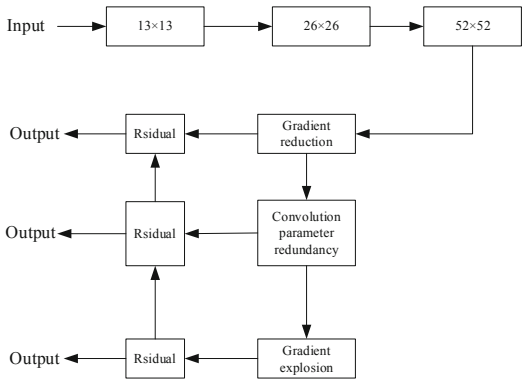

Fig. 1. Data input and output process

The overall structure of the improved algorithm is designed based on darknet-53 as the backbone network. In the network, 5 times of down sampling and the last 3 times of down sampling are used to get the feature map, as well as the prediction for three scales [6]. Finally, The feature pyramid is used to transfer the last two downsampled outputs to the previous one. Through up sampling and large-scale feature map, deep features and shallow features can be fused. Then, the residual network is used to output the three scales. Finally, through the non maximum suppression algorithm, the error and redundant prediction box is discarded, and the prediction box with accurate positioning and high confidence is retained as the final algorithm marker box. The comparison of input and output residuals is shown in Fig. 2.

The improved YOLO algorithm normalizes the input image to square 416 × 416 pixels, and then divides it into 8 × 8 grids. Each grid predicts 8 prediction boxes, and returns the confidence C and location information of each prediction box. The value of confidence C is calculated according to the following formula:

$$C = O_b \times IOU \frac{T}{pred} \tag{1}$$

Where, when the value of O_b is 1, it means there is an object, and its value is 0, indicating no object; IOU means the intersection and ratio between the prediction box and the actual region, and the prediction box returns five values of (x, y, w, h) and C values, so the last output tensor is $8 \times 8 \times (5 \times 8 + C)$. Category value is the value for each equal grid, and the category value C is for each prediction box. In the YOLO improved algorithm, the value of C is usually 7, so the image is normalized and divided

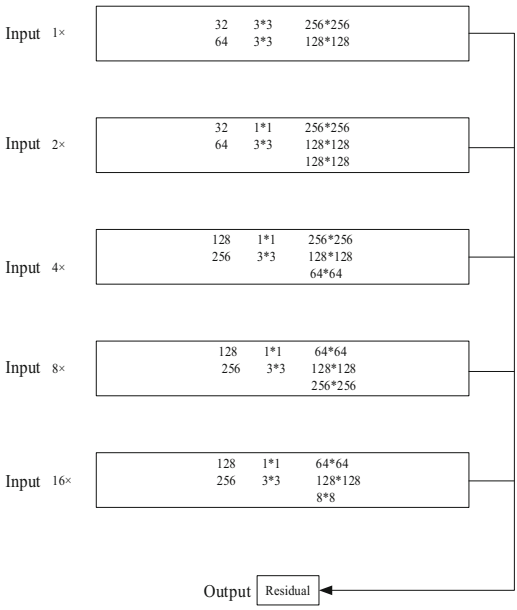

Fig. 2. Comparison of input-output residuals

into 7 × 7 mesh. O_b is usually 2, each grid is responsible for predicting two prediction boxes. In the YOLO improved algorithm, there are about 20 categories of data sets using the formula.

2.2 Noise Reduction for Abnormal Data Location

Because the shape of contour line is greatly affected by noise, a method of smoothing contour line is selected, that is, noise reduction. The improved YOLO algorithm is supported not by fitting the parameters of the contour line, but by fitting the regression contour line itself for smoothing, and the generalization error of the improved YOLO algorithm is small, which can avoid the phenomenon of "over fitting". Therefore, the improved YOLO algorithm is selected to smooth the nonlinear contour line, and the radial basis function is selected as the kernel function of the improved YOLO algorithm [7, 8].

When it is difficult to describe the functional form of a nonlinear contour with modeling method, a set of N contours with m points per contour can be regarded as an N × M matrix. A contour can be regarded as a point in high-dimensional space, so the contour data set can be regarded as n points in m-dimensional space. In the process of dimensionality reduction, the method of data depth can avoid the loss of data information in dimensionality reduction, and the data depth can judge the degree of data eccentricity when the data distribution is unknown or non normal distribution. Therefore, this paper selects the data depth to further process the smoothed data points. The improved YOLO algorithm is a common and effective data depth method. In this paper, the improved

YOLO algorithm is used to reduce the dimension of nonlinear contour data which is subject to normal distribution variation and non normal distribution variation.

In the YOLO improved algorithm, K-means clustering analysis can be used to classify data conveniently and efficiently. Therefore, the group of K-means is defined as 2 in the processing of identifying outliers. Assuming that outliers account for a small proportion in data, the data depth values transformed by YOLO algorithm are clustered and analyzed, and the few classes are distinguished from the data set as outliers, thus realizing the recognition of abnormal nonlinear contour [9]. Support vector regression has been widely used in various fields because of its small generalization error and can avoid the "over fitting" phenomenon. The basic operation idea is to find a real value function $f(x) = \frac{\omega x+b}{a}$ to minimize the expected risk. The expected risk can be expressed as follows:

$$R[f] = \int \frac{\omega x + b}{a} dP(x, y) \tag{2}$$

In the formula, $P(x, y)$ means to select the independent and identically distributed sample points to form the training set according to the probability distribution, and $\frac{\omega x+b}{a}$ means the loss function.

In this process, the support vector regression mechanism of the improved YOLO algorithm can be divided into linear and nonlinear regression models. Therefore, the nonlinear control problem is mainly studied. The basic principle of this method is to first map the original data set to Hilbert space through nonlinearity, and make the estimated regression function linear, and transform the training set into a linear one The following spatial coordinates:

$$\{(\phi(x_1y_1)), (\phi(x_2y_2)), \cdots, (\phi(x_3y_3))\} \tag{3}$$

Then, the approximate linear regression is made in this high-dimensional feature space. By introducing the penalty factor c, and using Lagrange function and KKT condition to calculate the kernel function, the quadratic programming problem is solved

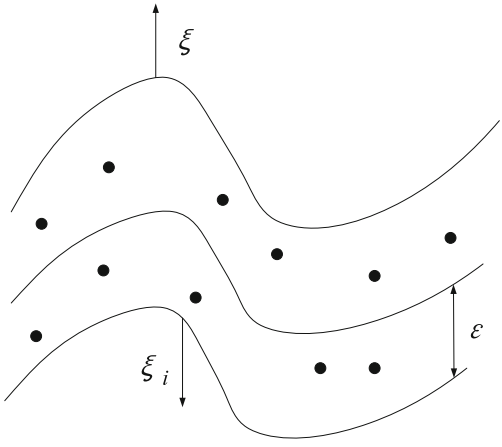

Fig. 3. Abnormal data positioning model

[10, 11]. Among them, the existence of relaxation variables ξ_i and ξ_i^* allows some data points to remain outside the confidence interval determined by c, as shown in Fig. 3, so as to obtain the optimized regression function.

Because the performance of Gaussian radial kernel function is better than other kinds of kernel functions, this paper uses RBF to check the nonlinear contour model for contour regression, so as to realize the abnormal data location based on the improved YOLO algorithm.

2.3 Design of Exception Location Data Recognition Based on Improved YOLO Algorithm

In the method of abnormal data recognition, neural network model training is mainly used to extract normal data samples and various types of abnormal data to establish the recognition model of abnormal data categories. In this process, the input information of the abnormal data identification model is mainly the load data collected by the collection system, and the output is the identification result. When it is normal data, the output result is the normal data type, and when it is abnormal data, the category of the abnormal data can be identified. In this paper, the abnormal data recognition method is mainly to predict the input data output through the improved YOLO algorithm calculation model completed by training. According to the output results, whether the input data is abnormal data can be judged, and the category of abnormal data can be judged [12].

Before using the recognition model, select the appropriate sample data, and optimize the weight attribute and threshold of the improved Yolo algorithm through genetic algorithm, and then use the sample data to train the improved YOLO algorithm. Therefore, the number and composition of the training sample data will have a great impact on the accuracy of the prediction output of the abnormal data recognition model, which plays a very important role effect. In general, the more sample data, the more knowledge the improved algorithm gets, the stronger its cognitive ability, and the more accurate the sample mapping. But in the current situation, due to the influence of environmental factors, it is impossible to obtain too much sample data. At the same time, too much training sample data not only can not improve the performance of the improved YOLO algorithm, but also can increase the network training time. The output accuracy of the abnormal data recognition model also depends on the parameter setting of the improved YOLO algorithm.

3 Experimental Research

3.1 Preparation of the Experiment

The whole experiment is implemented by Python language. IDE is eclipse pydev. All the data packages used in the process of data processing and algorithm implementation include numpy, pandas, Matplotlib, sciki learn, tensorflow, etc. Numpy is used to support a large number of dimensional array and matrix operations, and also provides a large number of mathematical function libraries, mainly for array operations, which belongs to an extension library of Python language. Pandas has been incorporated into a large

number of databases and some standard data models, which provides the tools needed to operate large data sets efficiently. The tool is mainly used to solve the data analysis task. Maplotlib is a 2D drawing library in Python, which generates publishing quality level graphics in various hard copy formats and cross platform interactive environments [13, 14]. Sciki learn is a machine learning algorithm library implemented by python, which is used to realize data preprocessing, classification, regression, dimension reduction, model selection and other common machine learning algorithms. Tensorflow is an open source software library which uses data flow chart for numerical calculation. The universality of this system makes it widely used in other computing fields.

In this simulation experiment, MATLAB software is used to calculate. Firstly, the nonlinear contour data which obey normal distribution variation and non normal distribution variation are generated by MATLAB simulation through Monte Carlo simulation. The data are identified by SDC method, and compared with the two conventional methods.

Firstly, 500 nonlinear contour lines are generated, including some proportion of abnormal contour, and the function relation of the nonlinear contour set meets the following requirements:

$$\begin{cases} Y = 3\cos(x) + 5\sin(x) + \varepsilon_N \\ Y = 3\cos(x) + 5\sin(x) + \varepsilon_A \end{cases} \tag{4}$$

In the formula, $x \in (0, 2\pi)$; ε_N is the noise factor of normal contour; ε_A is the noise factor of abnormal contour; In this experiment, the performance of the new method and the comparison method when the proportion of abnormal contour in the total is different and ε_N and ε_A take different distribution.

In this paper, the first kind of error, the second kind of error and the running time of the program are used as the evaluation indexes of the performance of each method to identify the abnormal contour type of the target. The first kind of error refers to the case that the normal control line is judged as the abnormal contour line, and its calculation formula is as follows:

$$\alpha = \frac{FN}{TP + FN} \tag{5}$$

In the formula, FN is the number of normal contour lines determined as abnormal contour lines; TP is the number of normal contour lines determined as normal contour lines; α is the probability of the first type of error. Similarly, the probability of the second error is obtained.

$$\beta = \frac{FP}{TN + FP} \tag{6}$$

In the formula, FP represents the number of abnormal contours determined as normal contours; TN represents the number of abnormal contours determined as abnormal contours; β represents the probability of the second type of error. Next, this experiment will verify the performance of the proposed method and the conventional method under the condition that the proportion of abnormal contour lines in the total is 0.1, 0.15, 0.2, and the target contour data is distributed differently.

3.2 Performance Test for Algorithm

Because the proportion of abnormal contour in the total may affect the value of the first kind of error and the second kind of error, this experiment explores the influence of the proportion of abnormal contour in the total of the first kind of error and the second kind of error while exploring the different values of mathematical expectation under the lognormal distribution of noise factor. Matlab is used for data simulation and classification. The parameter C of support vector regression is set to 1, and the parameter g is set to 0.0625. According to perc = 0.1, 0.15, 0.2, 500 contour lines are randomly generated each time. According to the different mathematical expectation values of noise factor lognormal distribution, each value is repeated 1000 times, that is, a total of 2000 times, and the first type error and the second type error are calculated incidence. The probability of the occurrence of a type of error is shown in Table 1 when the proportion of abnormal contour in the total is different.

Table 1. Type I error incidence

Mathematical expectations	The algorithm in this paper	Conventional algorithms 1	Conventional algorithms 2	Conventional algorithms 3
0.1	0.965	0.398	0.978	0.967
0.2	0.869	0.684	0.857	0.759
0.3	0.586	0.489	0.594	0.475
0.4	0.369	0.569	0.674	0.578
0.5	0.852	0.875	0.965	0.954
0.6	0.967	0.987	0.578	0.477
0.7	0.598	0.965	0.415	0.488
0.8	0.496	0.859	0.552	0.955
0.9	0.756	0.598	0.485	0.789
1.0	0.975	0.785	0.758	0.958

At the same time, under the influence of different regression parameters, the first contrast effect picture of error rate is obtained, as shown in Fig. 4.

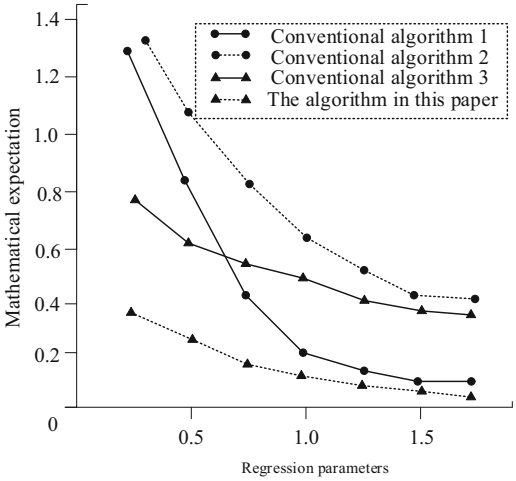

Fig. 4. Comparison of first error rate

As shown in Fig. 4, the image obtained by the algorithm in this paper is relatively flat, and has been given a relatively small value, because the error rate obtained by the algorithm is less affected during the change of regression parameters. The other three algorithms have very steep curves, especially the conventional algorithm 1 and the conventional algorithm 2. Therefore, this method has good performance in the first error rate test. The probability of the second type of error is shown in Table 2 when the proportion of abnormal contour is different.

Table 2. Type II error incidence

Mathematical expectations	The algorithm in this paper	Conventional algorithms 1	Conventional algorithms 2	Conventional algorithms 3
0.1	0.569	0.963	0.963	0.935
0.2	0.486	0.852	0.846	0.369
0.3	0.693	0.741	0.951	0.259
0.4	0.478	0.987	0.756	0.645
0.5	0.369	0.654	0.944	0.284
0.6	0.852	0.321	0.759	0.368
0.7	0.654	0.951	0.869	0.947
0.8	0.789	0.753	0.765	0.768
0.9	0.637	0.846	0.864	0.486
1.0	0.951	0.864	0.894	0.957

Under the influence of different regression parameters, the second effect chart of error rate comparison is obtained, as shown in Fig. 5.

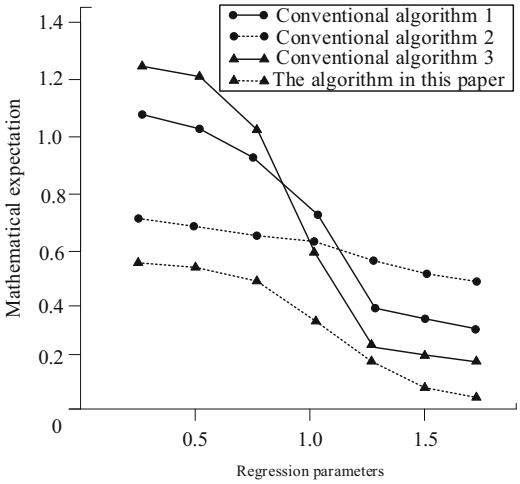

Fig. 5. Comparison of second error rate

As shown in Fig. 5, all four curves are of the "s" type structure with different slopes. The results show that the second error rate is least affected by the regression parameters. Among the other three algorithms, the slope of conventional algorithm 1 and conventional algorithm 3 is steepest, and the influence is the most. However, the conventional algorithm 2 has a relatively slow slope, but its value is larger. Therefore, according to the four curves, the method designed in this paper has the best performance in the second error rate test.

4 Conclusion

In order to improve the performance of the method in the case of non normal distribution, this paper proposes the comprehensive use of support vector regression, data depth and cluster analysis technology to identify the abnormal points of the first stage control chart of nonlinear contour. Then through the comparison between the new method and the conventional method, it is proved that this method has better data in the two evaluation indexes of the first type error rate and the second type error rate. The method of anomaly location data recognition is optimized. However, this study only identifies and detects for fixed datasets, and did not consider the characteristics of more data sets, such as nonstructural data. This will be deeply studied and analyzed in the next step.

References

1. Zhou, Z., Chen, Q., Ma, B., et al.: An improved YOLO target detection method with its application in cable device abnormal condition recognition. Electric. Measur. Instrum. **57**(2), 14–20 (2020)

2. Wang, L., Zheng, D.: anomaly identification of dam safety monitoring data based on convolutional neural network. J. Yangtze River Sci. Res. Inst. **38**(1), 72–77 (2021)
3. Wenli, J.I., Liutao, X.I., Bin, W.A.N.G.: Abnormal data recognition method of coal mine monitoring system based on imbalanced data set. Ind. Mine Autom. **46**(1), 18–25 (2020)
4. Xia, J., Liang, W., Wu, Z.: Research on automatic recognition algorithm of abnormal data in power monitoring based on mobile wavelet tree. Electron. Des. Eng. **28**(18), 148–152 (2020)
5. Zhang, H., Fan, Z., Chen, M. Application of isolated forest in abnormal identification of dam monitoring data. Yellow River **42**(8), 154–157, 168 (2020)
6. Lei, J., Chu, X., Jiang, Z., et al.: Abnormal automatic identification system data by visual analytics. J. Harbin Eng. Univ. **41**(6), 840–845 (2020)
7. Li, Y., Li, T.: Application of improved K-means algorithm in recognition of wind power abnormal data. Comput. Era **2**, 6–8 (2020)
8. Li, W.: Fast recognition and simulation of fuzzy anomaly data in nonlinear electronic networks. Comput. Simul. **36**(7), 351–354 (2019)
9. Xu, W.: Abnormal data recognition method based on power big data cleaning model. New Gener. Inf. Technol. **2**(17), 41–46 (2019)
10. Liu, S., Lu, M., Li, H., et al.: Prediction of gene expression patterns with generalized linear regression model. Front. Genet. **10**, 120 (2019)
11. Li, J., Zhang, R., Safonov, P., et al.: Outlier recognition method for spatio-temporal data based-on copula function and M-K test. Mod. Account. **39**(12), 3229–3236 (2019)
12. Xu, G., Hou, M., Xiong, H.: Moving target detection of remote tower based on improved YOLO algorithm. Sci. Technol. Eng. **19**(14), 377–383 (2019)
13. Liu, S., Liu, D., Srivastava, G., Połap, D., Woźniak, M.: Overview and methods of correlation filter algorithms in object tracking. Complex Intell. Syst. **7**(4), 1895–1917 (2020). https://doi.org/10.1007/s40747-020-00161-4
14. Fu, W., Liu, S., Srivastava, G.: Optimization of big data scheduling in social networks. Entropy **21**(9), 902 (2019)