Research on Unsupervised Anomaly Detection of Gas Heating Energy Consumption Based on Ensemble Learning

Lizhuo Gao, Huihua Yang*, Yanzhu Hu

Beijing University of Posts and Telecommunications, Beijing, China

Abstract

To improve the detection of abnormal data of heating energy of gas boilers, based on the gas heating energy consumption data of a place in Beijing in recent two years, this paper deeply studies and analyzes the detection effects, advantages and disadvantages of unsupervised anomaly detection algorithm, iForest, LOF and One-Class SVM algorithm models. Finally, based on the idea of ensemble learning, the weighted fusion of the above three algorithms is carried out, and the accuracy of anomaly detection in gas heating is successfully improved. The F1 value of the fusion model on the data set is about 95.8%.

Key words

Anomaly detection, Gas, Ensemble learning, Unsupervised

1. Introduction

There are abnormal conditions in the energy consumption of gas heating, such as natural gas pipeline leakage, furnace pipe scaling or corrosion, water shortage and pipe explosion. With the increase of the amount of data, there are more and more abnormal data, and it will become more and more difficult to extract effective information from the data. Therefore, it is necessary to improve the accuracy of abnormal detection with the help of machine learning method. At present, the research on abnormal detection of gas heating energy using machine learning algorithm is almost a piece of white paper. Machine learning methods are mainly divided into supervised learning and unsupervised learning. Supervised learning requires high-quality labels. These labels need to be manually labeled for the current data set, and then targeted training. Although the supervised detection effect is better, it has no generalization, so the supervised method is not desirable.

The algorithm mentioned in this paper are popular and unsupervised anomaly detection algorithms. Meanwhile, in other application fields of anomaly detection, many researchers have proposed integrated detection algorithms and achieved certain results. For example, Li Guocheng and others have proposed an isolated forest power theft detection algorithm based on bagging quadratic weighted integration, which improves the power theft detection effect[1]. Gao Xin et al. proposed an integrated algorithm for anomaly detection of power dispatching data based on log interval isolated forest[2]. The proposed method is progressiveness in the comprehensive performance of anomaly detection AUC value.

In summary, this paper adopts the unsupervised learning method to make an exploratory research on the abnormal detection of energy consumption of gas heating. Based on the idea of integration, iForest, LOF and One-Class SVM are weighted integrated as the base classifiers, and the fusion model has better effect on the gas data set.

ICCEIC2022@3rd International Conference on Computer Engineering and Intelligent Control

EMAIL: *1157951896@qq.com (Huihua Yang)

2. Anomaly detection model

2.1. Detection based on iForest algorithm

Forest (Isolated forest) algorithm detects outliers by isolating sample points, and isolates samples by using binary search tree structure of isolated tree[3]. In our given gas sampling point database, the vast majority of data points are surrounded together in space, reflecting similar characteristics, while outliers will be isolated from other data and isolated earlier.

2.1.1. Create iForest

The establishment of iForest mainly depends on the generation of iTree. The core steps are described below:

Step1: The root node was established and 100 samples were randomly selected as the sample set.

Step2: Specify the function variable and randomly select one important feature from the existing 12 parameters as the splitting basis of iTee.

Step3: Generates a random cutoff point in the selected maximum and minimum values of the specified variable.

Step4: The sample data segment is divided into two subspaces by extending a hyperplane from the cut-off point.

Step5: Determine whether the samples contained in each subspace are the same samples, or the iTree splitting times have reached log2n. If the conditions are met, splitting is terminated and an iTree is constructed. Otherwise proceed to Step3, Step4, and Step5.

According to the above steps, set the number of iTree N=50, and set the extraction method as repeatable sampling. The whole iForest can be established by repeating the above steps.

2.1.2. Calculate outlier score

100 randomly selected data samples can be evaluated by the iForest tree. When the data samples are spread all over iTree, the node depth of the samples in each tree is extracted, and the average depth of the samples in iForest is calculated. The abnormal score of the samples can be obtained by further transformation of the depth of the iTree. The formula for calculating abnormal scores is shown in (3):

$$
\begin{gathered}
s(x,n) = 2^{\frac{E(h(x))}{c(n)}} \\
C(n) = 2H(n-1) - \left(\frac{2(n-1)}{n}\right) \qquad (1) \\
H(k) = ln(k) + 0.577
\end{gathered}
$$

Where, h(x) is the depth of the node of sample x in iTree, E represents the average value, and c(n) represents the average length of the binary tree constructed by N sample points. In this application model, n=100. The closer s(x) score is to 1, the more likely the sample is to be an outlier, and the closer s(x) score is to 0, the more likely it is to be a normal sample.

2.2. Detection based on LOF algorithm

LOF(Local Outliers Factor) algorithm realizes anomaly detection by calculating the local density deviation of a given data point relative to its neighborhood [4]. We select the data at any time from a set of data as point p and calculate the local density of point p. The lower the density is, the more likely it is to be an outlier.

The k_{th} distance of the point p to be measured is defined as follows:

$$d_k(p) = d(p,o) \qquad (2)$$

And : (1) There are at least k points o, that do not include p in the set. (2) There are at most k-1 points o, that do not include p in the set. O, is the data point of other time in the sampling point database.

$N_k(p)$ is the k_{th} distance of p and all points including the k_{th} distance, so the number of k_{th} domain points of p is greater than or equal to k. Reach-distance $k(p,o)$ is defined as the reachable distance from o to p, which is at least the k_{th} distance of o, or the true distance of o, p.

The local outlier of point p is the average of the ratio of the local accessible density of the domain point $N_k(p)$ of point p to the local accessible density of point p. The significance of this formula is that if the ratio is closer to 1, the density difference between point p and field point is smaller, and point p and field point belong to the same class cluster. If this ratio is greater than 1, it indicates that the density of point p is less than that of its domain points, and p may be an outlier. Which can be expressed as:

$$\begin{aligned} lrd_k(p) &= \frac{|N_k(p)|}{\sum_{o\in N_k(p)} reachdist_k(p,o)} \\ LOF_k(p) &= \frac{\sum_{o\in N_k(p)}(lrd_k(o)/lrd_k(p))}{|N_k(p)|} \end{aligned} \tag{3}$$

2.3. Detection based on One-Class SVM algorithm

One-Class SVM is an unsupervised learning method, that is, it does not need us to mark the output label of the training set. There are many solutions for One-Class SVM to find the divided hyperplane and support vector machine[5]. Here is only a special idea SVDD. For SVDD, we expect that all samples that are not abnormal are positive categories. At the same time, it uses a hypersphere rather than a hyperplane to divide. The algorithm obtains the spherical boundary around the data in the feature space, hoping to minimize the volume of the hypersphere, so as to minimize the influence of abnormal point data.

Assuming that the parameters of the generated hypersphere are the center o and the corresponding hypersphere radius $r >$ 0, the volume $V(r)$ of the hypersphere is minimized, and the center o is a linear combination of support vectors; Similar to the traditional SVM method, the distance from all training data points x_i to the center can be strictly less than r. But at the same time, a relaxation variable with penalty coefficient C is constructed ζ_i. The optimization problem is shown in the following equation (4):

$$\begin{gathered} \underbrace{min}_{r,p} \sum_{i=1}^{m} \zeta_i \\ \|x_i - o\|_2 \leq r + \zeta_i, i = 1,2,3 \ldots m \\ \zeta_i \geq 0, i = 1,2,3 \ldots m \end{gathered} \tag{4}$$

After solving with Lagrange duality, we can judge whether the new data point Z is included. If the distance from Z to the center is less than or equal to radius r, it is not an abnormal point. If it is outside the hypersphere, it is an abnormal point.

2.4. Algorithm optimization

Anomaly detection can be regarded as a binary classification problem of anomaly and normal, so the three basic models are called base classifiers in this paper. Based on the idea of ensemble learning[6], this paper takes iForest, LOF and One-Class SVM as three base classifiers for ensemble learning.

The specific process is as follows:

If a sample point in the first base classifier has been accurately classified, its weight will be reduced when constructing the next classifier; On the contrary, if a sample point is not accurately classified, the weight is increased. Because there are many ways of weight distribution of training data, which will not be repeated here.

Then, the weight updated samples are used to train the next classifier, and the whole training process goes on like this. Finally, increase the weight of the base classifier with small classification error rate to make it play a greater decisive role in the final classification function, while reduce the weight of the base classifier with large classification error rate to make it play a smaller decisive role in the final classification function. In other words, the classifier with low error rate accounts for a large proportion in the final classifier, and vice versa (see **Figure 1**).

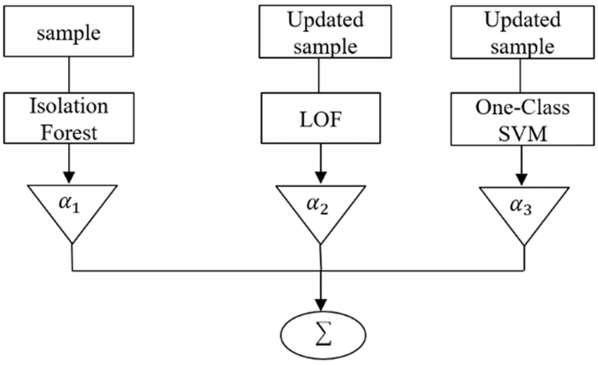

Figure 1 Isolated forest detection effect

Here is the calculation method of the weight coefficient of the base classifier, that is, first calculate the detection error rate of the above three anomaly detection models, then calculate the weight coefficient of the three base classifiers, and finally combine the three base classifiers.

Step1: Calculation error rate ε_t , as shown in the following equation (5).

$$\varepsilon_t = \frac{\sum_{i=1}^{N_t} I[h_t(x_i) \neq y_i]D_t(x_i)}{N_t} \tag{5}$$

Where $I[]$ is the discriminant function, Xi represents the ith sample in the training set, a total of n samples, t takes 1, 2 and 3 to represent the three basic classifiers of iForest, LOF and One-Class SVM respectively, $D_t(x_i)$ is the weight distribution of the training set of the t-th basic classifier, and $h_t(x_i)$ is the classification result of the i-th sample by the t-th basic classifier.

Step2：Calculate the weight α_t of base classifier h_t, as shown in the following equation (6).

Step3：Combine three base classifiers to get $f(x)$, as shown in the following equation (7). Thus, the final classifier is obtained, as shown in the following equation (8). Among them, three base classifiers are used, so T is 3.

$$\alpha_t\text{=log}\frac{1-\varepsilon_t}{\varepsilon_t} \tag{6}$$

$$f(x) = \sum_{t=1}^{T} \alpha_t h_t(x) \tag{7}$$

$$G(x) = sign(f(x)) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right) \tag{8}$$

3. Comparative experiment

3.1. Data Description

This paper collects the real heating data of boiler room a in Beijing from 2020 to 2022. The sample size after logarithmic fusion is 500000 and the dimension is more than 200. After expert selection and feature selection, the dimension is reduced to 53 dimensions. These data sets contain real time series, and abnormal data has also been marked. Some of the marks are artificially designed and added fault and alarm scenarios based on the experience of gas operation and maintenance experts. The abnormal proportion of the finally processed data sets is about 11%. The amount of data and labels can support the construction and evaluation of the model. Some of the real labels are shown in the **Table 1.**

Table 1 Partial label display

Time	Label
2022/1/4 9:30	Domestic water pump failure, abnormal return water in high and low areas and abnormal smoke exhaust
2022/1/4 9:40	Abnormal smoke exhaust of boiler
2022/1/4 9:50	Abnormal smoke exhaust of boiler
2022/1/4 10:00	Abnormal smoke exhaust of boiler
2022/1/4 10:10	Abnormal smoke exhaust of boiler

In the data preprocessing stage, the data with single point blank or only single point mutation of continuous variables are supplemented by the average values of the previous second and the next second. Discrete variables are directly filled with 0 value according to their meaning. At the same time, this paper also constructs the instantaneous heat released by gas and the heat absorbed by boiler water. On this basis, the change trend of heat proportion is estimated through calculation, which is used as the auxiliary index of the anomaly detection model in this paper. In feature selection, through SelectKBest in sklean, this paper uses the maximum information coefficient as the scoring function to select the features most related to the label.

3.2. Experimental Result

The parameter settings of iForest model are as follows: Max_samples= 120， contamination=0.11. The settings in the LOF model are as follows: n_neighbors=20, leaf_size=30, contamination=0.11. Euclidean distance measurement is adopted. One-Class SVM adopts Gaussian kernel function, and the parameter gamma of RBF kernel type is set to 0.1. The contamination represents the proportion of outliers in the total data volume.

The performance of anomaly detection using iForest, LOF or One-Class SVM is very general. In fact, this is because each algorithm is good at dealing with different scenes and characteristics, as follows: LOF believes that the outliers are the ones with large deviation from the local density of a given data point relative to its neighborhood. In other words, it pays more attention to the local; iForest believes that in the data space, the sparsely distributed areas mean that the probability of data occurring in this area is very low, so it can be considered that the data falling in these areas is abnormal. In other words, it pays more attention to the whole; One-Class SVM is an algorithm for anomaly detection based on the characteristics of normal data. It considers all data with similar characteristics to normal data as normal data, otherwise it is considered as abnormal data. However, in the complex scenario of energy consumption of gas heating, the manifestations of anomalies are complex. For example, the abnormal data of water pipe scaling is very similar to the normal data. In other words, the three models have their places that are not well considered. Therefore, based on the idea of integration, this paper takes iForest, LOF and One-Class SVM as base classifiers, combines them with reference to the weighted method of AdaBoost, complements their advantages, and forms an anomaly monitoring model with better performance in gas heating energy consumption detection(see **Figure 2).**

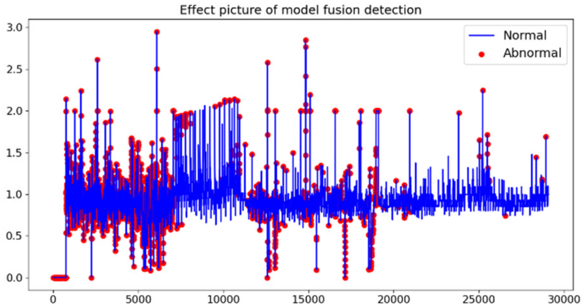

Figure 2 Fusion model inspection renderings

3.3. Evaluation

The anomaly detection algorithm finally divides the data into normal data and abnormal data, which is a binary classification problem. Therefore, the anomaly detection algorithm uses the evaluation standard F-score value of the classification model to evaluate(see **Table 2)**.

Table 2 Imapact assement

Method	Evaluating Indicator			
	P	R	AUC	F1
iForest	0.6387	0.6687	0.9321	0.6587
LOF	0.8199	0.9567	0.9874	0.8810
One-Class SVM	0.8043	0.9999	0.9882	0.8915
Fusion model	0.9274	0.9958	0.9745	0.9579

It can be seen that the accuracy and recall rate of the model are very high. Among them, the recall rate of One-class SVM model is very high, but the accuracy is low. Overall, the F1 value of the improved model is about 0.96, and achieved good results.

4. Conclusion

Based on the idea of integration, this paper establishes an unsupervised anomaly detection model of gas heating energy consumption through weak classifier integration, and compares it with other popular anomaly detection models. This paper mainly makes two contributions: first, unsupervised anomaly detection model is applied to the heating energy anomaly detection of gas boiler room for the first time, and relevant exploratory research is carried out in this field, which has a strong reference value for the application of gas heating anomaly detection combined with machine learning. Second, the detection effect is improved by integrating the fusion model, and the F1 value reaches 0.5 on the existing data set 0.9579, which is 2 to 1 percentage points higher than the single anomaly detection model. Of course, when this model is applied to different gas fired boilers in practical application, there are problems such as insufficient universality, long training time and high complexity, which will be further improved in the subsequent research.

5. References

[1] Li Guocheng. : An isolated forest power theft detection algorithm based on Bagging quadratic weighted integration [J]. Power System Automation, 2022, 46 (2): 92-100.

[2] Gao Xin. : A method of power dispatching data anomaly detection based on log interval isolation [J] Power grid technology, 2021, 45 (12): 10

[3] He Kun. : Study on identification of electrical anomalies in isolated forest based on PCA[J]. Computing Technology and Automation, 2021, 40(2) : 76-80

[4] Zhang Shuo. : Outlier detection algorithm based on grid LOF and adaptive K-means [J]. Command Information Systems and Technology, 2019, 10(1): 90-94

[5] Li Chengliang. : Study on anomaly detection method of Airborne Tacang Ranging Information based on One-Class SVM [J]. Modern navigation, 2015(3): 282-285, 309

[6] Li Yuan. : Research on fault detection based on K-means clustering and local outlier factor algorithm [J]. Chemical automation and instrumentation, 2019, 46(10): 816-821