Anomaly identification model of gas boiler heating energy based on anomaly detection algorithm

Dan Liu1*, DongHong Huang2, Xinyu Ye3, XinLi Dong4
1BEIJING GAS GROUP CO., LTO., 100035 Beijing, China
2 BEIJING GAS GROUP CO., LTO., 100035 Beijing, China
3 BEIJING GAS GROUP CO., LTO., 100035 Beijing, China
4 BEIJING GAS GROUP CO., LTO., 100035 Beijing, China
*Corresponding author’s e-mail: liudan02@bjgas.com

Abstract. As an efficient and clean energy in China's energy structure, the proportion of natural gas in the energy structure is increasing. Because the data of gas boilers are very complex, there is no obvious relationship, and the data changes with time, so there is some difficulty in detecting data anomalies. Therefore, in this paper, isolated forest algorithm, One-Class SVM algorithm and LOF algorithm are used to build a model to detect the outlier of multidimensional time series and capture the abnormal fluctuations in the data. Through model training on the data set of gas group, it is found that the isolated forest algorithm, one-class algorithm and LOF algorithm all have good function of distinguishing outliers, but in this data set, the isolated forest algorithm has better effect, and the identification of outliers conforms to the theory of energy conservation model.

1. Introduction
With the continuous development of China's urbanization and the implementation of the strategy of changing coal to gas, as a "lifeline" energy, the scale of users is constantly expanding, which provides favourable help for the country's economic transformation and development. The energy consumption of heating in China accounts for a large proportion in the energy consumption structure of the whole country, the heating energy consumption is huge, and the efficiency is low, and the energy consumption per unit area is about several times that of developed countries, so from the point of view of energy consumption, there is a broad space for the energy-saving development of China's heating industry.

Through the analysis of a large number of heating data, we can well find the heating law in the heating system, find the abnormality in the heating, and plan and formulate a reasonable heating temperature; it can provide data support for the actual heating, so as to analyse the causes of energy waste, save energy more reasonably, and achieve the purpose of energy saving.

2. The main content of this article
Because of the characteristics of large data, many types, large increment and fast increment of gas-fired boiler data, the deep-seated reasons behind abnormal data can not be extracted by using traditional statistical methods. At the same time, the calculation efficiency of anomaly detection algorithm is low. In order to improve the accuracy and efficiency of model detection, in this paper,
isolated forest algorithm, One-Class SVM algorithm and LOF algorithm are used to build a model to
detect outliers in multidimensional time series and capture abnormal fluctuations in data.

3. Research status
Now, domestic and foreign experts have put forward different solutions to the abnormal analysis of
gas boiler time series data, including support vector machine, neural network, clustering analysis, local
outlier factor method, statistical method and so on.

A.J.FOX introduced the idea of statistical diagnosis into time series analysis in 1972, and after
defining the outliers of time series, there have been many studies on outlier monitoring of time series
at home and abroad. Junshui Ma proposes to use single pattern SVM to detect abnormal patterns of
time series, that is, by transforming the data of time series into high-dimensional space, abnormal
patterns that deviate from the data can be detected by training[1]. Deng Yujie and Zhu Qingsheng of
Chongqing University put forward the method of abnormal data mining based on clustering. In 2008,
Liu et al proposed an integrated learning algorithm—iForest[2]. IForest is different from other
anomaly detection methods. The main idea of iForest is to distinguish normal points and outliers by
using the feature that outlier data is easy to be "isolated". The earliest concept of LOF was proposed
by Breuning et al[3]. On the basis of LOF, some people proposed the COF[4] algorithm, which judged
the outliers with the average connection distance between the data instance and its neighborhood as
the outliers score. Ester et al. proposed DBSCAN[5] algorithm, through which many clusters could be
found. Zhang et al proposed the LDOF[6] algorithm, which used the relative position to determine the
distance between the data instance and its nearest neighbor, so as to judge the outliers.

4. Technical scheme

4.1. Isolated forest anomaly detection
Isolated forest algorithm is a fast anomaly detection method based on set, which belongs to
unsupervised learning algorithm, that is, it does not need to define parameter model and history
training samples, and has linear time complexity and high accuracy.

For the data of gas-fired boiler, the lower depth indicates that it is farther than the normal data, and
the probability of abnormal data is higher. Conversely, a higher depth indicates that the data is closer
and more likely to be normal data than normal data.

The way of isolating forest division data can be expressed by isolated tree. The isolated tree is a
bipartite tree, the leaf node is data, and the non-leaf node is represented as a partition. Each partition
randomly obtains the characteristic \( Q \) and the feature partition value \( p \) of the data, and divides the data
into two parts according to \( p \). Then divide the segmented data again.

Suppose a gas boiler dataset with sample size \( n \), after the tree is built by isolated forest algorithm,
the average depth of the tree is shown in Formula 1, where \( H(i) \) is the harmonic number, this value
can be estimated as

\[
c(n) = 2H(n - 1) - 2(n - 1)/n
\]  

(1)

Suppose \( h(x) \) is the depth of data sample \( x \) in the isolated tree. \( E(h(x)) \) is the average depth of the
isolated forest of data sample \( x \), where the value of \( E(h(x)) \) ranges from 0 to \( \text{nnurl} \). The isolated
value can be calculated by calculating the depth of the average data and test data in the isolated tree, as
shown in Formula 2:

\[
s(x, n) = 2 \frac{E(h(x))}{c(n)}
\]  

(2)
4.2. One-Class SVM anomaly detection

SVM is a new machine learning method based on the VC dimension theory of statistical learning theory and the principle of structural risk minimization. Its basic idea is to find an optimal classification hyperplane to separate the two types of samples as much as possible under the condition of ensuring high classification accuracy and maximizing the classification interval, so that the classifier can improve its generalization ability. That is, the confidence interval is minimized by maximizing the classification interval, and the higher the classification accuracy is, the smaller the empirical risk is. Scholkopf et al. extended the two-classification problem of SVM to a class of classification problems, so they proposed a class of support vector machines (One-class Support vector Machine, referred to as One-Class SVM). In essence, One-Class SVM is an unsupervised learning classification method.

Given a set of training samples \( T = \{x_1, x_2, \ldots, x_l\}, x_i \in \mathbb{R}^d \), since the sample points need to be projected into the high-dimensional feature space, the kernel function must still be introduced, and then a hyperplane must be found to separate the training sample points from the origin by the maximum interval. The following secondary planning problems need to be solved:

\[
\begin{align*}
\min_{\omega, \rho, \xi} & \frac{1}{2} \|\omega\|^2 + \frac{1}{v} \sum_{i=1}^{l} \xi_i - \rho \\
\text{s.t.} & (\omega \cdot \phi(x_i)) \geq \rho - \xi_i \\
& \xi_i \geq 0, i = 1, \ldots, l
\end{align*}
\]

(3)

Among them, \( \xi_i \) is a relaxation variable, and its function is to punish some of the misdivided sample points. \( v \in (0, 1] \) controls the distance from the aggregation region of the sample points to the maximum interval hyperplane and the distance from the origin to the maximum interval hyperplane.

Similarly, the formula (3) is transformed into the optimization problem of dual variables by Lagrange multiplier method. Then the following Lagrangian function can be constructed to solve the problem:

\[
L(\omega, \rho, \xi, \alpha, \gamma) = \frac{1}{2} \|\omega\|^2 + \frac{1}{v} \sum_{i=1}^{l} \xi_i - \rho - \sum_{i=1}^{l} \alpha_i \left[ (\omega \cdot \phi(x_i)) - \rho + \xi_i - \gamma_i \right] - \sum_{i=1}^{l} \gamma_i \xi_i
\]

(4)

\( \alpha \) and \( \gamma \) represent Lagrangian factors.

The next step is to solve the partial derivative of each parameter, as follows:

\[
\frac{\partial L}{\partial \omega} = \omega - \sum_{i=1}^{l} \alpha_i \phi(x_i) = 0
\]

(5)

\[
\frac{\partial L}{\partial b} = -1 + \sum_{i=1}^{l} \alpha_i = 0
\]

(6)

\[
\frac{\partial L}{\partial \xi_i} = \frac{1}{v} - \alpha_i - \gamma_i = 0
\]

(7)

By substituting formula (5), formula (6) and formula (7) into formula (4), the dual problem of formula (3) can be obtained:
According to the KKT condition, the optimal solution $\alpha^*$ needs to meet the following conditions:

$$\alpha^*_i \left[ (\omega \cdot \phi(x_i)) - \rho + \xi_i^* \right] = 0$$  \hspace{1cm} (9)

$$\left( \frac{1}{vl} - \alpha^*_i \right) \xi_i^* = 0$$  \hspace{1cm} (10)

Further, according to the satisfied support vector, we can obtain:

$$\rho^* = \sum_{i=1}^{l} \alpha_i^* K(x_i, x_j)$$  \hspace{1cm} (11)

Finally, the decision function of One-Class SVM is:

$$f(x) = \text{sign} \left( \omega^* \cdot \phi(x) - \rho^* \right)$$

$$= \text{sign} \left[ \sum_{i=1}^{l} \alpha_i^* K(x_i, x) - \rho^* \right]$$  \hspace{1cm} (12)

For a given test sample point, the value of $f(x_i)$ can be obtained from the above decision function, and then the category of the sample point can be distinguished according to the following conclusion:

1. When $f(x_i) > 0$, then the test sample point $x_i$ belongs to the positive sample, that is, the target sample point;

2. When $f(x_i) < 0$, then the test sample point $x_i$ belongs to the negative sample, that is, the abnormal sample point.

4.3. LOF (local factor detection)

LOF algorithm is a classical algorithm in density-based local anomaly data mining algorithm. This method measures the anomaly degree of data objects by calculating the local anomaly factor LOF (Local Outlier Factor), of each object in the data set. If the local anomaly factor is larger, the degree that the data is an abnormal data object is greater, and vice versa. The density-based local anomaly factor LOF (LocalOutlier Factor) algorithm mainly involves the calculation of k-distance, k-distance neighborhood, reachable distance and reachable density of data objects. Fig.1 vividly explains the k distance of p and the neighbor object of p.
As shown in Fig.1 above, there are six \( p_1, p_2, p_3, p_4, p_5, p_6 \) data objects in the Fig.1, in which the distance between the data objects \( p_2, p_3, p_4, p_5, p_6 \) and \( p_1 \) is indicated. Suppose there are 1 data objects in the circle and \( n-k-1 \) data objects outside the circle, where \( n \) represents the number of data objects in the data set, then the distance between the data objects \( p_4 \) and \( p_1 \) on the circumferential boundary is the \( k \) distance of \( p_1 \). The data object on the circumferential boundary and the data object in the circle is the neighbor object of \( p \), namely \( \{p_2, p_3, p_4\} \).

Define 1 \( k \)-distance: point out the maximum distance to its \( k \) nearest neighbor, which can be expressed as \( k-distance(P) \).

Define 2 \( k \)-distance neighborhood: the neighborhood of \( k \)-distance of a point is defined as a set of all points in the space area with the center of the point and the radius of \( k-distance(P) \). That is:

\[
N_{k-distance}(P) = \{Q \mid P \in D \land d(P, Q) \leq k - \text{tan ce}(P)\} 
\]  

(13)

Among them, point \( Q \) is called the \( k \) nearest neighbor object of point \( P \).

Define 3 Reachable distance: given a natural number \( k \), the reachable distance of point \( P \) to point \( O \) is:

\[
reach-dist_k(P, Q) = \text{Max}\{k - \text{tan ce(}O), dist(P, O)\} 
\]  

(14)

Define 4 local reachable density: the local reachable density of any given point is based on the reciprocal of the average reachable density of the \( k \) nearest neighbor of \( P \). That is,

\[
\text{lrd}_k(P) = \frac{1}{\sum_{O \in N_{k-distance}(P)} \text{reach-dist}_k(P, Q)} 
\]  

(15)

Where \( N_{k-distance}(P) \) is a collection of objects in the \( k \) neighborhood of point \( P \). When the sum of reachable distances of \( k \)-nearest neighbor objects is 0, the local reachable density of points is infinite.

Define 5 The local exception factor \( \text{LOF} \) is defined as:
The local outlier factor of point P quantifies its deviation level, which is obtained by calculating the mean value of the local reachable density ratio of point P to its k nearest neighbor. It is obvious that the smaller the local reachable density of the lattice is, and the higher the local density of the k-nearest neighbor of the lattice is, the higher the LOF value of the point is.

5. experimental result

Operating environment of the hardware: the computer CPU is Pentium(R) Dual-core T4300@2.10GHz, the memory is 2.0GB, and the hard disk is 320G. Software running environment: Windows 10, VC++6.0

The experiment adopted the gas boiler data set of Beijing Gas Group, which contained 400,000 data objects. The Fig.2 below shows the outliers trained by the three models.

![Anomaly detection model](image1)

**Fig.2** Anomaly detection model

![The heat ratio of the model](image2)

**Fig.3** The heat ratio of the model

It can be seen from the three-dimensional view in model renderings 3 that the isolated forest model, One-Class SVM model and LOF algorithm have a good performance in distinguishing the outliers from the normal ones. It can be seen from the renderings of heat ratio in Fig.3 that the outliers are all distributed at the low or high positions with sharp changes. It indicates that the data fluctuated greatly during this period, which was probably due to the error of sensor recording, resulting in the large fluctuation of heat proportion, which was identified as an outlier by the model.

From the above three model rendering shows that One-Class SVM algorithms identify anomalies are mainly based on distance from the Angle of the data, this model mainly for a drastic change, whether the data still is not sensitive to local anomalies point problem, so the model for the gas boiler in terms of data set, we use the effect not beautiful. The LOF local anomaly factor algorithm is...
sensitive to large data fluctuations, but compared with the isolated forest, the model performed poorly in this data set. Compared with One-Class SVM, the isolated forest has the best performance in this data set for the identification of outlier in accordance with the energy conservation model theory.

6. conclusion
In this paper, based on the gas group gas boiler data set, isolated forest algorithm, One-Class SVM algorithm and LOF algorithm are used to build a model to detect the outlier of multidimensional time series, capture the abnormal fluctuations in the data, and establish the relationship between gas energy and boiler energy consumption.

Acknowledgments
This work was financially supported by the scientific research project of BEIJING GAS GROUP CO., LTO., (20jk013)

Reference
[1] Aasim and Singh S.N. and Mohapatra Abheejeet. Data driven day-ahead electrical load forecasting through repeated wavelet transform assisted SVM model[J]. Applied Soft Computing Journal, 2021, 111.
[2] Wang Zhenrui, Zhao Kunyu, CAI Chuan, Ding Mengzhen, Wang Peng. Abnormal ship behavior analysis based on DBSCAN and iForest algorithm [J]. Ship electronic engineering,2021,41(04):89-94.
[3] Geng Juncheng, Guo Zhimin, Li Xiaolei, Su Juan, Yuan Shaoguang, Niu Shuangxia. LOF and SVM based power distribution network cable variable relationship data verification method [J]. The Chinese test,2021,47(04):49-54.
[4] Qiu Qing-gong. Outlier Mining algorithm based on clustering outlier factor and unique nearest neighbor set [D]. Yanshan university,2019.
[5] de Moura Ventorim Igor et al. BIRCHSCAN: A sampling method for applying DBSCAN to large datasets[J]. Expert Systems With Applications, 2021, 184.
[6] Hu Tian-yu, Liu Song. Research on intrusion detection technology based on Chi-square test and LDOF algorithm [J]. Journal of Qilu University of Technology,2019,33(03):62-69.