High-Dimensional Data Anomaly Detection Framework Based on Feature Extraction of Elastic Network

Yang Shen[1], Jue Bo[1], KeXin Li[2(✉)], Shuo Chen[1], Lin Qiao[1], and Jing Li[2]

[1] State Grid Liaoning Electric Power Supply Co., Ltd., Shenyang 110004, Liaoning, China

[2] College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, Jiangsu, China

kexinli@nuaa.edu.cn

Abstract. Although appropriate feature extraction can improve the performance of anomaly detection, it is a challenging task due to the complex interaction between features, the mixture of irrelevant features and relevant features, and the unavailability of data tags. When conventional anomaly detection methods deal with the problem of anomaly detection of high dimensional data, the performance of anomaly detection will be degraded due to the existence of irrelevant features. This paper proposed a method of feature extraction and anomaly detection for high dimensional data based on elastic network, which can filter irrelevant features and improve the accuracy and efficiency of anomaly detection. In this paper, an outlier scoring method was used to score the outliers of the original data, and then outliers and the original data were input into the elastic network for sparse regression. After feature extraction of elastic network, those irrelevant features to abnormal data are ignored, thus reducing the dimension of data. Finally, high-dimensional data are detected efficiently according to extracted features. In the experimental stage, we used the high-dimensional anomaly dataset provided by ODDS to detect the performance of the proposed method based on AUC detection accuracy, ROC curve, feature number, convergence speed and other indicators. The results show that the proposed method not only can effectively extract the features related to high-dimensional anomaly data, but also the detection accuracy of outliers has been greatly improved.

Keywords: Elastic network · Feature extraction · Anomaly detection · High-dimensional data · Data mining

1 Introduction

An outlier is a data point that is significantly different from the remaining data. Hawkins defined [1] an outlier as “An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different

X. B. Zhai et al. (Eds.): MLICOM 2019, LNICST 294, pp. 3–17, 2019.
https://doi.org/10.1007/978-3-030-32388-2_1

mechanism." Outliers are also referred to as abnormalities, discordant, deviants, or anomalies in the data mining and statistics literature. In most applications, the data is created by one or more generating process, which could either reflect behaves unusually, it results in the creation of outliers. Therefore, an outlier often contains useful information about abnormal characteristics of the systems and entities that impact the data generation process. The recognition of such unusual characteristics provides useful application-specific insights. Some examples are as follows:

Intrusion detection systems [2]: In many computer systems, different types of data were collected about the operation system calls, network traffic, or other user actions. This data may show unusual behavior because of malicious activity. The recognition of such activity was referred to as intrusion detection.

Credit-card fraud [3]: Credit-card fraud had become increasingly prevalent because of greater ease with which sensitive information such as a credit-card number can be compromised. In many cased, unauthorized use of a credit card may show different patterns, such as buying sprees from particular locations or very larger transactions. Such patterns can be used to detect outliers in credit-card transaction data.

Power equipment failure: the effective analysis of distribution network data can not only meet the needs of planning and design, production scheduling, load forecasting, power quality and power decision-making, but also solve the problems faced by the future distribution network, such as accurate energy supply, power demand side management and decentralized energy storage. With the deep integration of sensor measurement, information communication, analysis and decision-making technology and modern distribution network data, massive heterogeneous, polymorphic and high-dimensional load data are produced. Abnormal data in these data can often reflect the operation status of power equipment. If abnormal data can be detected in time, the status of power equipment can be timely diagnosed, thus avoiding unnecessary losses.

The so-called high-dimensional data refers to the data with a high number of dimensions, which can reach hundreds of thousands or even higher. There are two main difficulties in analyzing and processing high-dimensional data. First, the problem that Euclidean distance cannot be used. In low dimensional space, Euclidean distance can be used to measure the similarity between data, but in high dimensional space, distance does not make much sense. The second is the disaster of dimensionality. With the increasing dimensionality, many of the conventional outlier detection methods do not work very effectively. This is an artifact of the well-known curse of dimensionality. In high-dimensional space, the data becomes sparse, and the true outliers become masked by the noise effects of multiple irrelevant dimensions, when analyzed in full dimensionality. Besides, the amount of calculations for analyzing and processing these data will increase rapidly, and the cost of calculation will increase exponentially. Therefore, in the process of abnormal data detection for high-dimensional data, the following challenges are encountered: (i) High-dimensional data usually contain features unrelated to abnormal data. These extraneous features will affect the anomaly detection of high-dimensional data. (ii) As the increase of data dimension, relevant concepts in low-dimensional space such as neighbor, distance and nearest neighbor will be unusable, resulting in the inability to use conventional abnormal data detection methods based on distance, density and so on. Conventional abnormal data detection methods tend to use all the features of the data in the process of abnormal data detection, but these data often contained many useless features, which will lead to a large deviation in the results of abnormal data detection. In addition, there may be redundancy between different features to reduce the efficiency of anomaly detection. At present, there are many methods for detecting abnormal data, such as probabilistic statistical model-based method [4, 5], linear model-based method [6–8], and nearest-neighbor based method [9–11]. However, due to the computational complexity and efficiency of these methods, it will take a lot of cost to carry out anomaly detection on high-dimensional data, and they do not perform particularly well in the aspect of anomaly detection effect of high-dimensional data. Therefore, these methods cannot be simply applied to the anomaly detection of high-dimensional data, which needs to be processed and then detected by these methods.

Filtering irrelevant features can significantly improve the detection effect and performance of abnormal data detection, but it is more difficult to extract and remove irrelevant features when there is a complex interaction between features. At present, there is little work on feature extraction in abnormal data detection. Most feature extractions were devoted to classification, regression and clustering [12–14]. In terms of outlier feature extraction, most of the existing work focuses on unbalanced classification [15] and data categories, without considering how to filter the data used in the abnormal data detection process. Elastic network [16] is a sparse regression model. In the process of sparse regression, the purpose of feature extraction is achieved by continuously shrinking the coefficients of the relevant features. If there is a correlation between features, elastic network will select two or more features from them, so as to ensure that the selected features are the most representative and correct. In this paper, elastic network was used to extract features from high-dimensional data so as to filter irrelevant features to reduce the impact of irrelevant features on abnormal data detection. A high-dimensional anomaly data detection model based on multi-level feature extraction is constructed by elastic network. The model firstly used the existing anomaly detection approach score the original data, and then the abnormal scores and the original data were input to the elastic network so as to extract the features which were most related to the abnormal data. Then the extracted features were used to score the abnormal data again, and the above steps are continuously cycled until the loss function of elastic network will no longer reduced. Finally, the abnormal score obtained by the above process is integrated to obtain the final detection result. According to the work done in this paper, the main contributions are as follows:

(i) A feature extraction method of high-dimensional data based on elastic network was proposed. In the process of extracting features of abnormal data by elastic network, the outlier score was used as target feature and the original data was used as predictors to filter irrelevant features of high-dimensional data.

(ii) A cyclic feature extraction and anomaly data detection model was constructed. In the process of detection abnormal data, we use isolated forest algorithm [17] to achieve outlier scoring of data, and then use elastic network and outlier score to extract features of high-dimensional data. According to the loss function of elastic network, the above process is looped until the loss function of elastic network is no longer reduced, so as to realize the loop feature extraction and anomaly detection of abnormal data.

(iii) A multi-level detection method for anomaly data based on elastic network was proposed. At each level, different thresholds were set for the abnormal score calculated by the isolated forest method, so as to select different data input into the elastic network, and then extract more feature related to abnormal data. Finally, the extracted features and detection results of each level are integrated.

2 Related Work

Abnormal data detection [18] is a very popular research direction in the field of data mining. The general problem of identifying outliers has been addressed by very different methods that can be roughly classified as global versus local outlier models. A global outlier model [5] leads to a binary decision of whether or not a given object is an outlier. A local outlier method rather assigns a degree of outlierness to each object. Such an "outlier factor" is a value characterizing each object in "how much" this object is an outlier. Many classical approaches for abnormal data detection, such as probabilistic and statistical models [2] for outlier detection, linear models for outlier detection [6], proximity-based outlier detection [8], etc. were used to process low dimensional data. The earliest methods for outlier detection were rooted in probabilistic and statistical models and data back to the nineteenth century. These methods were proposed well before the advent and popularization of computer technology and were therefore designed without much focus on practical issues such as data representation or computational efficiency. Nevertheless, the underlying mathematical models are extremely useful and have eventually been adapted to a variety of computational scenarios. The main assumption in linear models is that the normal data is embedded in a lower-dimensional subspace. Data points that do not naturally fit this embedding model are, therefore regarded as outlier. In the case of proximity-based methods, the goal is to determine specific regions of the space in which outlier points behave very differently from other points. On the other hand, in linear methods, the goal is to find lower-dimensional subspaces, in which the outlier points behave very differently from other points. This can be viewed as an orthogonal points of view to clustering- or nearest-neighbor methods, which try to summarize the data horizontally (i.e., on the rows or data values) rather than vertically (i.e., on the columns or dimensions). Proximity-based techniques define a data point as an outlier when its locality (or proximity) is sparsely populated. The proximity of a data point may be defined in a variety of ways, which are subtly different from one another but are similar enough. The most common ways of defining proximity for outlier analysis are as follows:

Distance-Based [7]: The distance of a data point to its k-nearest neighbor (or other variant) is used in order to define proximity. Data points with large k-nearest neighbor distance are defined as outliers. Distance-based algorithms typically perform the analysis at a much more detailed granularity that the other two methods. On the other hand, this greater granularity often comes at a significant computational cost.

Density-Based [8]: The number of other points within a specified local region (grid region or distance-based region) of a data point, is used in order to define local density.

These local density values may be converted into outlier scores. On kernel-based methods or statistical methods for density estimation may also be used. The major difference between clustering and density-based methods is that clustering methods partition the data points, whereas density-based methods partition the data space.

Cluster-Based [9]: The non-membership of a data point in any of the clusters, its distance from other clusters, the size of the closest, or a combination of these factors are used to quantify the outlier score. The clustering problem has a complementary relationship to outlier detection problem in which points either belong to cluster or they should be considered outliers.

Clearly, all these techniques are closely related because they are based on some notion of proximity (or similarity). The major difference is at the detailed level of how this proximity is defined. These different ways of defining outliers may have different advantages and disadvantages. In many cases, the distinctions between these different classes of methods become blurred when the outlier scores are defined using more than one of these concepts. Many real data sets are very high dimensional. In some scenarios, real data sets may contain hundreds of thousands of dimensions. With increasing dimensionality, many of the conventional outlier detection methods do not work very effectively. This is an artifact of the well-known curse of dimensionality. In high-dimensional space, the data becomes sparse, and the true outliers become masked by the noise effects of multiple irrelevant dimensions, when analyzed in full dimensionality. A main cause of the dimensionality curse is the difficulty in defining the relevant locality of a point in the high-dimensional case. For example, proximity-based methods define locality with the use of distance functions on all the dimensions. On the other hand, all the dimensions may not be relevant for a specific test point, which also affects the quality of the underlying distance functions. For example, all pairs of points are almost equidistant in high-dimensional space. The challenges arising from the dimensionality curse are not specific to outlier detection. It is well known that many problems such as clustering and similarity search experience qualitative challenges with increasing dimensionality. The subspace-based outlier detection method can process high-dimensional data. The more successful method [19] is to identify multiple related subspaces for candidate anomaly data, and then combine the results from different subspaces to create a more robust collection-based ranking. While many collection methods for subspace analysis have achieved great success, the particularly difficult case is that a large number of dimensions are weakly related (but not very relevant), and even more dimensions are not locally relevant.

3 The Proposed Approach

Aiming at the problem that the abnormal data detection of high-dimensional is faced with hundreds of thousands dimensions and irrelevant features, this paper proposed a cyclic feature extraction method based on elastic network to filter irrelevant features, so as to reduce the dimension of data and realize the detection of abnormal data in a lower data space. The method included three parts: isolated forest anomaly score calculation function, anomaly selection function based on Chebyshev theorem and elastic network feature extraction. In step t of the method, for a given data set $X = \{X_1, X_1, \cdots, X_N\}$, each data has M features, namely $X_i = (x_1, x_2, \cdots, x_M)$ and their abnormal score vector $S^{t-1} \epsilon R^N$. The abnormal selection function $H(S^t)$ is used to select a part of the data as the possible abnormal data C^t, and then the selected abnormal data and abnormal scores S^t are input into the elastic network feature selection module ElN^t to obtain the mean square error (mse^{t} for short) and the features F^t which are most relevant to the abnormal score. Then, the selected features were used to perform the abnormal score by the isolated forest abnormal score calculation function G. Repeat the above steps until the mean square error of the current elastic network is greater than the previous mean square error, i.e. $mse^{t+1} > mse^t$ or the current detection accuracy is less than the initial detection accuracy, i.e. $AUC(S^t) < AUC(S^0)$. The overall framework of the method is shown in Fig. 1.

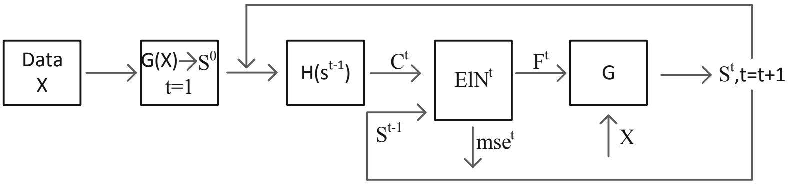

Fig. 1. The framework of feature extraction and anomaly detection by elastic network

Among them, the G is a kind of abnormal scoring function, we used isolated forest to calculate the abnormal score in this paper. H is a selection function, and some possible abnormal data is selected according to the abnormal score S^t. ElN^t is elastic network, which mainly performs feature extraction of high-dimensional data. When the above loop is completed, we integrate the sequence of abnormal score to obtain the final abnormal scores, so as to realize the anomaly detection of high-dimensional data.

3.1 The Isolated Forest Abnormal Score Calculation Function *G*

Isolated forest uses the idea of random sampling to randomly sample the original data set and then constructs the isolated tree on the sub-samples. This method considers that the data close to the root node is more likely to be abnormal data. The process of constructing the isolation tree, calculating the path length and calculating the abnormal score is as follows:

iTree: The construction method of isolation tree is similar to the construction process of binary tree. Suppose T is a node of the isolation tree, then T is either an external node with no child nodes or an internal node with two children nodes. On each sub-sample randomly select a data attribute as separate property, and put all the data in the root node, and then randomly selected a data as the segmentation value (the value between maximum and the minimum) from the selected attribute. The value about selected attribute of the data less than the segmentation value was divided into left child node, otherwise was divided into right child node. Do the same for the left child and right child until all the data is on the external node or at the height specified in the isolation tree.

Path Length: The path length $h(x)$ of the data record x refers to the number of edges x traverses an iTree from the root node until the traversal is terminated at an external node.

Outlier Score: After constructing the isolation tree, calculate the path length of the data x in different isolation trees, then obtain the average value and normalize the abnormal score by the normalization factor. The specific calculation method is:

$$E(h(x)) = \sum\nolimits_{i=1}^{n} h_i(x)/n \tag{1}$$

$$s(x, \varphi) = 2^{-\frac{E(h(x))}{c(\varphi)}} \tag{2}$$

Where, $s(x, \varphi)$ represent the outlier score; $h_i(x)$ represents the path length of the data x in the i-th iTree; $E(h(x))$ represents the average path length of the data x in all the iTree; $c(\varphi)$ denotes the normalization factor, which means the maximum path length in the corresponding iTree, and $c(\varphi)$ is calculated as:

$$c(\varphi) = \begin{cases} 2H(\varphi - 1) - \frac{2(\varphi-1)}{\varphi}, & if\ \varphi > 2 \\ 1, & \varphi = 2 \\ 0, & otherwise. \end{cases} \tag{3}$$

Where, φ is sample size, $H(\varphi)$ is the harmonic function that can be calculated by $ln(\varphi) + 0.5772156649$ (Euler constant).

3.2 Anomaly Selection Function Based on Chebyshev's Theorem $H(S^t)$

After getting the abnormal score of the data, we think that the abnormal score is a random variable X, and then we calculate the expectation and variance of X. Suppose the expectation of X is $E(X) = \mu$ and the variance is $D(X) = \sigma^2$. According to Chebyshev's inequality, we can get:

$$P(X - \mu \geq \varepsilon) = \begin{cases} \leq \frac{\sigma^2}{\sigma^2 + \varepsilon^2}, & if\ \varepsilon > 0 \\ \geq 1 - \frac{\sigma^2}{\sigma^2 + \varepsilon^2}, & else \end{cases} \tag{4}$$

In the above formula, taking $\varepsilon = \alpha\sigma$, we can get:

$$P(X \geq \mu + \alpha\sigma) = \begin{cases} \leq \frac{1}{1+\alpha^2}, & if\ \varepsilon > 0 \\ \geq 1 - \frac{1}{1+\alpha^2}, & else \end{cases} \tag{5}$$

Through the above formula, we obtain the range of values corresponding to the event $|X \geq \mu + \alpha\sigma|$. In this paper we considered that the X corresponding to the event $|X \geq \mu + \alpha\sigma|$ is a possible abnormal data. The anomaly data is usually different from the distribution of most data or data that is significantly deviated from most data objects, and is only a small part of the entire data set. Therefore, we control the number of elements in the candidate set of outliers by the value of α. After obtaining the outlier scores S^t at different stages, the average value μ and the variance of the abnormal scores are calculated as σ^2, and the outliers are composed of data satisfying the following conditions:

$$C = \{(X_i, S_i)|H(S_i, \alpha) \geq 0\}, \forall X_i \in X, S_i \in S \tag{6}$$

Where $H(S_i, \alpha) = S_i - \mu - \alpha\sigma$, the value of α is 1.732 in the loop (i.e., the upper bound for false positive in η is 25%).

3.3 Elastic Network Feature Extraction Function ELN^t

Elastic network is a linear regression model that combines lasso regression and ridge regression and uses L_1 and L_2 norm as prior regular terms for training. The combination of lasso and ridge regression allows the learning of a sparse model with a small number of non-zero parameters, thereby ignoring those features with zero parameters to realize feature selection. Elastic network can realize automatic selection of variables and continuous contraction at the same time. When there is a high correlation between multiple variables or features, elastic network will select two or more of them, so as to preserve as many features as possible while selecting features. In addition, based on the lasso regression, the elastic network draws on the idea of ridge regression to achieve the double contraction of the correlation coefficient, which makes the elastic network inherit the stability of the ridge regression. In this paper, the process of feature extraction of high-dimensional data using elastic network is based on the possible anomaly data C^t. The process takes the abnormal score as the target feature and the original feature of the data as predictor, so as to find those relevant feature about the abnormal score. The sparse regression principle of elastic network for high-dimensional data is:

$$EIN(C, \lambda) = argmin_{\omega}\left(\frac{1}{2N}\sum\nolimits_{i=1}^{N} \left(S_i - X_i^T\omega\right)^2 + \lambda * a\|\omega\|_1 + \frac{a(1-\lambda)}{2}\|\omega\|_2\right) \tag{7}$$

Among them, N is the number of possible abnormal data records, ω is the regression coefficient, λ is a non-negative regularization parameter, and the value of a determines the relationship between elastic network and lasso regression and ridge regression. With the gradual increase of the parameter λ, the number of non-zero coefficients in the regression coefficient obtained by the Eq. (7) will gradually decrease, thus achieving the purpose of feature extraction. After obtaining the coefficient of regression, we can get the feature most relevant to the outlier score by (8), and realize the feature extraction:

$$F = \{X_i|\omega_i \neq 0, 1 < i < M\} \tag{8}$$

The selection of the parameter λ in an elastic network is crucial, and an inappropriate λ can lead to over-fitting or under-fitting. For this problem, we use 10 cross-validation methods on the possible outlier C^t to select the optimal λ to minimize the mean square error.

When the loop elastic network execution ends, it is assumed that a total of T times are performed. In this process, a series of abnormal scores $S^t \epsilon R^N, 1 \leq t \leq T$ and mean square error $mse^t, 1 \leq t \leq T$ are generated.

(1) Firstly sum the mean square error mse^t, and get $SUM = \sum_{i=1}^{T} mse^i$;
(2) Subtract SUM from each mean square error mse^t to get a new error term $MSE^t, 1 \leq t \leq T$ i.e. $MSE^t = SUM - mse^t$;
(3) Normalizing the MSE^t to obtain a series of weights $\omega^t, 1 \leq t \leq T$, i.e.

$$\omega^t = \frac{MSE^t}{\sum_{i=1}^{T} MSE^i}, 1 \leq t \leq T \tag{9}$$

(4) Unitize the outlier score S^{t}, i.e.

$$\tau^t = \frac{S^t}{||S^t||}, 1 \leq t \leq T \tag{10}$$

(5) Calculate the final outlier score.

$$S\_L = \frac{1}{T}\sum_{t=1}^{T} \omega^t \tau^t \tag{11}$$

4 Multi-level Feature Extraction and Outlier Detection Model Based on Elastic Network

In the process of feature extraction and anomaly detection of high-dimensional data using elastic network, the value of α in the anomaly data selection function $H(S^t)$ based on Chebyshev's theorem is fixed. The fixed value of α limited the range of feature extraction and the effect of anomaly detection to a certain extent. Here, we extend the above high-dimensional data anomaly detection method based on elastic network loop feature extraction to build a multi-level feature extraction and anomaly detection model. In each layer, the value of α can be adjusted according to the detection results of the previous level. If the detection result is too small of upper level, adjusting the value of α in this layer appropriately, and then using elastic network for feature extraction and anomaly detection again. The specific framework is shown in the Fig. 2:

In the above model, the α value of the lower layer is appropriately adjusted according to the detection result of the upper layer, thereby adjusting the feature extraction and abnormality detection of the layer. After obtaining the detection result of each level $S\_L_{\mathrm{i}}, 1 \leq \mathrm{i} \leq \mathrm{n}$, the results of each level are combined and averaged to obtain the final abnormality detection result.

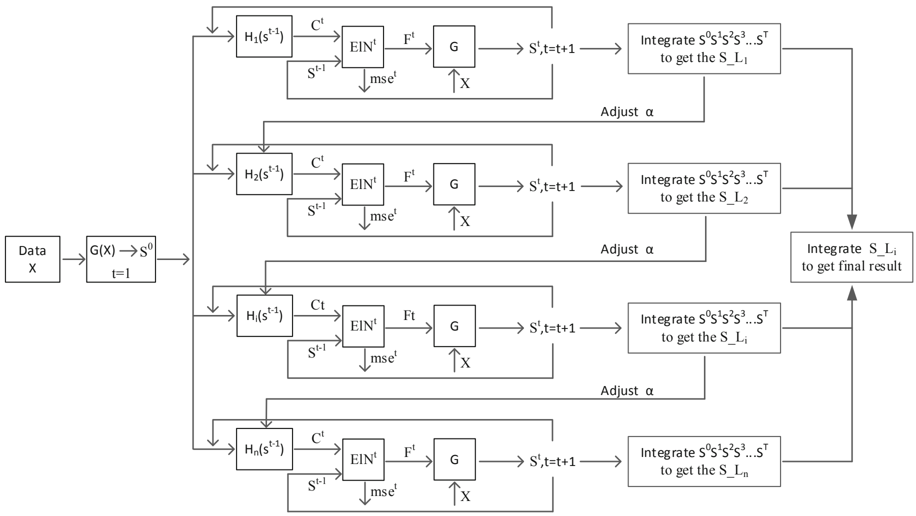

Fig. 2. The framework of multi-level elastic network feature extraction and outlier detection

5 Experiment

In order to verify the effectiveness of the proposed method, we selected 12 of the detection data sets of high-dimensional abnormal data provided by ODDS and built the experimental environment based on MATLAB 2018b, 64-bit Win7 operate system, Inter (R) Core (TM) i5-4460 3.20 GHz and etc. The method in this paper is compared with the methods of LOF, CARE and iForest by using ROC curve, detection accuracy AUC, extracted feature number, elastic network convergence speed and other indicators. In the LOF algorithm, the number of nearest neighbors K is set to 6. The number of isolated trees in the iForest method is set to 100, and the sampling size φ is set to 256. The specific information of the data set used in the experiment is as follows (Table 1):

(1) Extracting feature numbers

Noise or extraneous features are usually included in high-dimensional data, which are often ignored by conventional outlier detection methods. The high-dimensional abnormal data detection model based on elastic network firstly extracted the features of high-dimensional data in the process of anomaly detection. The characteristics of elastic network sparse regression were used to filter the noise irrelevant to the abnormal score, so as to screen out the features most related to the abnormal score, and then the high-dimensional data was abnormal detected by using these selected features. The feature extraction of the elastic network in this paper is to establish the correlation between the original data features and the anomaly scores, and then carry out sparse regression for the corresponding coefficients. In the procedure of regression, while keeping the objective function unchanged, the coefficients of corresponding features are constantly reduced until the coefficient of some features is 0, so as to achieve the filtering of irrelevant features. When there is a correlation between the original features,

Table 1. The information of experiment dataset

Dataset	Instance	Dimension	Outlier (%)
AID362	4279	117	60 (1.4%)
aPascal	12695	64	176 (1.3%)
Annthyroid	6832	6	534 (7.42%)
Arrhythmia	452	274	66 (15%)
Pima	768	8	268 (35%)
Cardio	1831	21	176 (9.6%)
L-Recognition	1600	32	100 (6.25)
Mnist	7603	100	700 (9.2%)
Musk	3062	166	97 (3.2%)
Optdigits	5216	64	150 (3%)
Satellite	6435	36	2036 (32%)
Sateimage-2	5803	36	71 (1.2%)

Table 2. Extracted feature number

Dataset	Original feature number	Extract feature number
AID362	117	64
aPascal	64	46
Annthyroid	6	3
Arrhythmia	274	131
Pima	8	4
Cardio	21	9
L-Recognition	32	18
Mnist	100	58
Musk	166	48
Optdigits	64	28
Satellite	36	23
Sateimage-2	36	22

the elastic network will select two or more of them to retain the features related to the abnormal score as much as possible. In addition, elastic networks retain the stability of ridge regression and can always produce effective solutions. In this paper, elastic network is used to extract the feature number of different data sets, as shown in Table 2:

Elastic network realizes feature extraction of high dimension data by shrinking the relevant regression coefficient while keeping the training error as small as possible. When the parameter λ in the elastic network is increasing, the coefficient of the feature will gradually decrease until it reaches 0, thus achieving the purpose of feature extraction. Here we show the contraction process of different feature coefficients in the process of feature extraction by elastic network. Each curve in the graph represents trajectory of each independent variable (feature), where the abscissa is the value of λ,

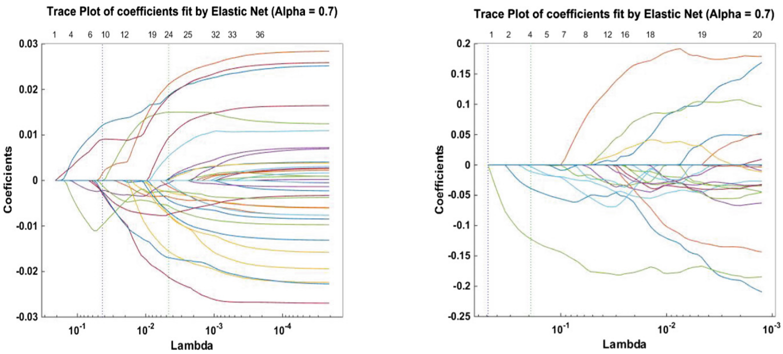

Fig. 3. The contraction process of coefficients during feature extraction

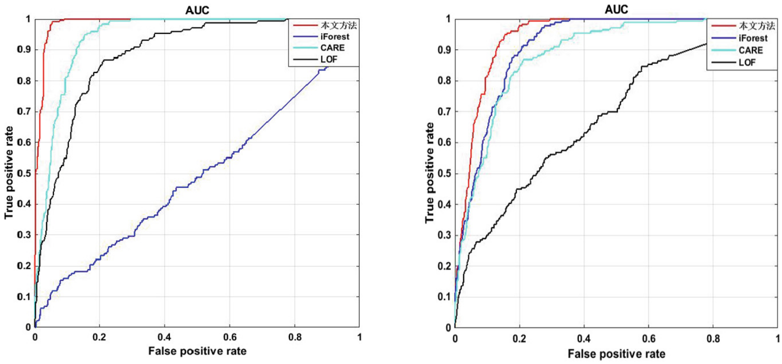

Fig. 4. The ROC curve on Cardio and Mnist dataset

and the ordinate is the size of the coefficient. We can see that the coefficient of the independent variable is decreasing to zero with the increase of the λ (Fig. 3).

(2) Detection accuracy

The ROC curve and the value of AUC can well reflect the detection accuracy. Generally, the larger of the area under the ROC curve (i.e., AUC), the higher the detection accuracy will be. Intuitively speaking, if one ROC curve can completely cover another ROC curve, the detection accuracy of the covered curve will be lower than the other one. We compared the detection accuracy of different abnormal data detection algorithm by the ROC curves and AUC values in the experimental data set. For the ROC curve, we selected the Cardio and Mnist datasets as representatives, and the results from the other datasets were compared by the area under the ROC curve (AUC) (Fig. 4).

It can be seen from the above ROC curve that the ROC curve of the proposed method in the data set Cardio and Mnist can completely cover the ROC curve of other methods, which indicates that the proposed method is superior to other methods in terms of detection accuracy. In order to fully demonstrate that the effect of the method

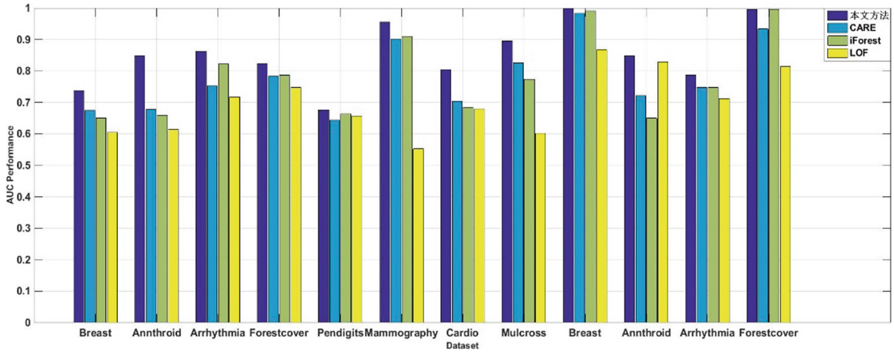

Fig. 5. The comparison of AUC of different outlier detection methods

is better than other methods compared with the method in this paper, we have compared it on other data sets. The results are shown in the Fig. 5. We use AUC as an indicator to measure the accuracy of detection. It can be seen from the AUC histogram that the detection accuracy of the proposed method is improved on several other data sets.

(3) Convergence rate of elastic network

In this paper, the anomaly detection of high-dimensional data is realized through the looping execution of isolated forest anomaly scoring function, the anomaly selection function based on Chebyshev theorem and the feature extraction of elastic network. In the feature extraction process of the elastic network, the value of λ was selected by cross validation because the improper selection of λ would lead to over-fitting or under-fitting of the elastic network. The green circle and dotted line locate the λ with minimum cross-validation error. The blue circle and dotted line locate the point with minimum cross-validation error plus one standard deviation. In the selection of the elastic network parameter λ, we select the green circle and dotted line located λ to ensure the minimum cross-validation dislocation. The change of training error of elastic network with λ is shown in the Fig. 6:

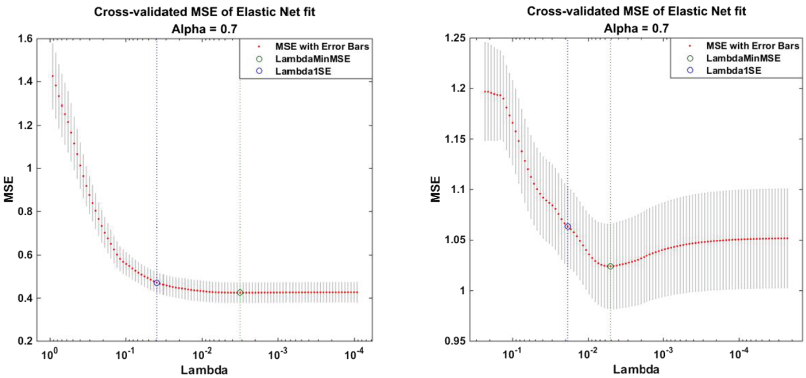

Fig. 6. The training error of the elastic network varies with lambda

6 Conclusion

In order to solve the problem of large dimension and low precision in anomaly detection of high-dimensional data, this paper proposed a method of anomaly detection of high-dimensional data based on elastic network feature extraction, and designed a multi-level anomaly detection model. The method consisted of three parts: the isolated forest anomaly score calculation function, the function to select the anomaly based on Chebyshev theorem and the function of feature extraction based on elastic network. The abnormal score and the outlier data selected by the outlier selection function are taken as the input of the elastic network, and the sparse regression is carried out so as to achieve the purpose of feature extraction. In this paper, the loop execution of the above three functions is used to achieve continuous feature extraction, and the abnormal score generated in the above process is integrated to obtain the final abnormal score, so as to realize the anomaly detection of high-dimensional data. The experimental results show that the proposed method can effectively realize feature extraction in the process of anomaly detection of high-dimensional data, and the accuracy and efficiency of anomaly detection were improved.

Acknowledgement. The research was supported by the State Grid Liaoning Electric Power Supply CO., LTD, and we are grateful for the financial support for the "Key Technology and Application Research of the Self-Service Grid Big Data Governance (SGLNXT00YJJS1 800110)".

References

1. Hawkins, D.M.: Identification of Outliers, 1st edn, pp. 11–12. Springer, Berlin (1980). https://doi.org/10.1007/978-94-015-3994-4
2. Li, L., Hao, Z., Peng, H., et al.: Nearest neighbors based density peaks approach to intrusion detection. Chaos Solitons Fractals **110**, 33–40 (2018)
3. Wang, C., Dong, H.: Credit card fraud forecasting model based on clustering analysis and integrated support vector machine. Cluster Comput. **16**, 1–6 (2018)
4. Gao, X., Fang, Y.: Penalized weighted least squares for outlier detection and robust regression (2016)
5. Chen, T., Martin, E., Montague, G.: Robust probabilistic PCA with missing data and contribution analysis for outlier detection. Comput. Stat. Data Anal. **53**(10), 3706–3716 (2009)
6. Aggarwal, C.C.: Linear models for outlier detection. In: Aggarwal, C.C. (ed.) Outlier Analysis, pp. 75–99. Springer, New York (2013). https://doi.org/10.1007/978-1-4614-6396-2_3
7. Dalatu, P.I., Fitrianto, A., Mustapha, A.: A comparative study of linear and nonlinear regression models for outlier detection (2016)
8. Pérez, B., Molina, I., Peña, D.: Outlier detection and robust estimation in linear regression models with fixed group effects. J. Stat. Comput. Simul. **84**(12), 2652–2669 (2014)
9. Xu, H., Mao, R., Liao, H., et al.: Closest neighbors excluded outlier detection. In: Online Analysis & Computing Science, pp. 105–110. IEEE (2016)

10. Liu, J., Wang, G.: Outlier detection based on local minima density. In: IEEE Information Technology, Networking, Electronic & Automation Control Conference, pp. 718–723. IEEE (2016)
11. Vwema, P., Yadava, R.D.S.: Fuzzy c-means clustering based outlier detection for SAW electronic nose. In: Convergence in Technology, pp. 513–519. IEEE (2017)
12. Bolón-Canedo, V., Sánchez-Maro, N., Alonso-Betanzos, A.: Recent advances and emerging challenges of feature selection in the context of big data. Knowl.-Based Syst. **86**, 33–45 (2015)
13. Du, L., Shen, Y.D.: Unsupervised feature selection with adaptive structure learning (2015)
14. Yu, K., Wu, X., Ding, W., et al.: Scalable and accurate online feature selection for big data. ACM Trans. Knowl. Discovery Data **11**, 16 (2015)
15. Zhang, C., Wang, G., Zhou, Y., et al.: Feature selection for high dimensional imbalanced class data based on F-measure optimization. In: International Conference on Security. IEEE (2018)
16. Feng, S., Sen, P.N.: Percolation on elastic networks: new exponent and threshold. Phys. Rev. Lett. **52**(3), 216–219 (1984)
17. Liu, F., Ting, K.M., Zhou, Z.H.: Isolation-based anomaly detection. ACM Trans. Knowl. Discovery Data **6**(1), 1556–4681 (2012)
18. Yang, Z., Zheng, Y., Gao, Y., et al.: Abnormal data detection for an e-business using object-oriented approach. In: Integration and Innovation Orient to E-Society, vol. 1 (2007)
19. Zimek, A., Gaudet, M., Campello, R.J.G.B., et al.: Subsampling for efficient and effective unsupervised outlier detection ensembles. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM (2013)