Anomaly Detection Method for Operation and Maintenance Data Based on One-Class Learning

Junhong Duan\textsuperscript{1,a}, Bo Yang\textsuperscript{1,b}, Xiang Wei\textsuperscript{1,c}, Ning Liu\textsuperscript{1,d}, Yintie Zhang\textsuperscript{2,e}

\textsuperscript{1}State Grid Gansu Electric Power Company Information and Communication Company. Lanzhou, China

\textsuperscript{2}Information System Integration Branch of NARI Technology Co., Ltd. Nanjing, China

\textsuperscript{a}duanjh@gs.sgcc.com.cn, \textsuperscript{b}yangbo@gs.sgcc.com.cn, \textsuperscript{c}weixiang@gs.sgcc.com.cn, \textsuperscript{d}liuning@gs.sgcc.com.cn, \textsuperscript{e}Zhangyintie@sgepri.sgcc.com.cn

Abstract. Due to frequent changes in system services, the amount of data collected in a certain state is insufficient, which causes problems such as insufficiency of normal sample data, scarcity of fault sample data, and lack of prior knowledge. Aiming at the problem of anomaly detection of small sample operation and maintenance data lacking negative samples, this paper proposes an operation and maintenance data anomaly detection method based on one-class learning, which uses SVDD (support vector data description) method to eliminate abnormal data in the collected operation and maintenance data. Then, we can better analyze the subsequent data. The experiments show that the proposed method is reasonable and effective.

1. Introduction

With the wide and deepening of IT applications, more and more business needs require the support of IT systems, and each enterprise attaches great importance to the maintenance and management of computer hardware and software resources. A large number of operation and maintenance technology tools and operation and maintenance management tools record the generated data at a time, and each information system runs a large amount of process data. The performance monitoring data generated by system operation and maintenance, as well as the log files of various systems, constitute the company's huge information system operation and maintenance data. The analysis of operation and maintenance data is undoubtedly providing an effective basis for operation and maintenance personnel to discover system problems and handle problems. It can help the operation and maintenance personnel to find out the cause of abnormal operation of the system or prevent the occurrence of system failure to some extent, which can turn passive maintenance into active prevention. Then the focus of operation and maintenance personnel is shifted to risk analysis and fault handling, it can effectively avoid risks, improve the efficiency of operation and maintenance personnel, and improve the fault monitoring and risk warning capabilities of operation and maintenance systems.

In the process of analyzing operation and maintenance data, it is necessary to cull the collected abnormal data firstly. Since the state of the business system changes rapidly, there are not many samples collected in a certain state. In addition, this paper assumes that the abnormal data collected is insufficient when the system is running stably, so it is necessary to adopt an anomaly detection method.
based on one-class learning. The anomaly detection method based on one-class learning only trains the data of normal classes, and the solution is different from the traditional two-class classification problem. This method can be used to solve the problem of missing samples due to the rare occurrence of abnormal data and the high cost of abnormal data acquisition. Therefore, this paper selects the Support Vector Data Description (SVDD) algorithm for scenarios with small samples and lack of abnormal data.

2. Related work

2.1. Data anomaly detection method

In many cases, anomaly detection is used to distinguish between normal and fault states. From the perspective of pattern recognition, this is a two-category problem. The traditional pattern recognition method designs a binary classifier by training two types of sample data. In some cases, there is an extreme problem that it is almost impossible to obtain multiple types of samples or it cost high to obtain samples, that is, only one type sample can be used to train the classifier, since the classification boundary decision of the two-class method requires the support of two types of example samples. When a certain type of sample cannot provide good decision support for some reason, the classification surface will be seriously deviated, so the method based on the two-class classifiers cannot adapt well to this classification problem, such problems need learn from the sampled target class samples and form a description of the data for the category, and then judge the abnormality of the new sample according to the given rules, that is, an anomaly detection method based on the one-class learning is needed.

2.2. Anomaly detection method based on one-class learning

At present, there are many solutions to the problem of anomaly detection and one-class classification. According to the principle of classification method, these methods can be roughly divided into three categories: density estimation method, reconstruction method and support boundary method [1].

(1)Density estimation method: The density-based method is not designed for a one-class problem, but a general method of pattern recognition. It estimates the density model of the training sample by a given one-class training sample, and sets a threshold. Above the threshold is normal data, below the threshold is abnormal data. Representative algorithms are: Gaussian model, mixed Gaussian model, Parzen window method [4] and k-nearest neighbor algorithm [4]. The disadvantage of the Gaussian model is that it can only describe the distribution of single-mode data. The model is too simple and lacks sufficient flexibility [2]. The mixed Gaussian model can model the general data distribution [2, 3]. As a further approximation to the real distribution, this model mainly approximates the multi-peak distribution of data by artificially setting the number of Gauss distributions, so its density function is a linear combination of various Gauss components. The main problem of this method is that the number of mixed components is data-dependent, so it is difficult to choose. Moreover, the mixture of multiple Gauss models requires more samples to overcome the dimension disaster. Parzen window method uses Parzen window estimation to estimate density, usually using Gauss kernel function. Because density estimation requires a large number of training samples, this method has the problems of low test efficiency and dimension disaster, and is only suitable for low-dimensional data. The core idea of the k-nearest neighbor algorithm is that if the majority of the k most neighboring samples in a feature space belong to a certain category, the sample also belongs to this category and has the characteristics of the samples on this category. The method determines the category to which the sample to be classified belongs based on only the category of the nearest one or several samples in determining the classification decision. This method has limitations in parameter selection and is sensitive to noise.

(2)Reconstruction method: When the density estimation method is not feasible, a simple model can be reconstructed by mapping the target data as an approximation to the original target data distribution model. The reconstruction-based method generates a detection model by training a given one-class sample. The reconstruction-based method generates a detection model by training a given single-class sample. It is assumed that the target class sample fully complies with the model, and the mapping of
the target class sample in the model space is used for the detection of new data. The new data is used as the abnormality criterion of the data by using the reconstruction error mapped by the detection model. Representative algorithms are: k-means [6] method, Principal Component Analysis (PCA) technology and neural network-based methods. The k-means method is mainly based on the fact that the samples are sufficiently aggregated and can be represented by a series of central points (mean), and the distance thresholds set to these centers are used as reconstruction errors for abnormality detection. The disadvantage of the method is that the selection of the number k of clusters is an open problem, which is heavily dependent on the initialization of the cluster center and is sensitive to the outliers. The principal component analysis technique is used to perform linear dimensionality reduction, and then the reconstruction error is calculated. The method obtains a set of orthogonal feature vectors represented by data in the least mean square sense by linear combination of data features to capture the maximum change direction of the data, and the feature vector corresponding to the large feature values is the main component [7]. On the reconstruction of nonlinear space, the method based on kernel principal component analysis can realize anomaly detection by calculating and minimizing the reconstruction error in high-dimensional space. In addition, neural network methods have been widely used, such as Autoassociators [9], AutoEncoders [10], and Self-organizing Map (SOM) [11, 12]. Based on the mature neural network research, this method inherits the self-organization and self-adaptive ability of the neural network to carry out training with only a little prior knowledge. There is a good classification effect on large-scale and non-linear practical problems, but the disadvantage is that there is only one type of sample for learning. The improved algorithm inevitably faces various threshold choices, such as training the neural network to determine the number of hidden layers of the network and the number of neurons per layer. The value of the number cannot be specified in advance, and only a trial method can be used to determine a reasonable network size.

3) Support boundary method: Support Vector Machine (SVM) is a learning method based on statistical learning theory of small sample learning [13]. This method can effectively solve high dimensional data modeling problems under finite sample conditions. At the same time, it also has the advantages of dimension insensitivity, convergence to global optimization, and generalization ability. The support boundary method for one-class classification draws on the maximum interval theory of SVM. By learning the target sample, it obtains a boundary or region around the target class, such as hyperplane, super (ellipsoid) sphere and so on. Then, the high-density area in the target data is mapped into a closed area to be separated from the non-target data. On the basis of a given empirical error, the correct reception rate of a given normal sample is ensured by minimizing the volume while minimizing the false acceptance rate of the abnormal sample. Representative methods are: One-Class SVM (OCSVM) based on hyperplane boundary segmentation [14] and Support Vector Data Description (SVDD) based on hypersphere boundary segmentation [15] and related improved method. The method based on support boundary embodies the fact that the one-class classifier focuses on describing the essence of the target data, and is suitable for dealing with one-class classification problems with small samples, high-dimensional and noisy data. Because of the intuitive description of the data, the use of kernel techniques makes it easier to solve in high-dimensional feature space. Therefore, the method based on support boundary becomes the most popular single-class classification method.

3. Operation and maintenance data abnormal deviation detection method

3.1. Support Vector Data Description (SVDD, Support Vector Data Description)
This paper uses SVDD as a single-class classification algorithm to help eliminate abnormal samples in the acquired operation and maintenance data. SVDD [16] is a support vector data description, first proposed by Tax et al. The basic idea is to find a hypersphere surrounding the target sample point in the feature space mapped to the high dimension. Through minimizing the volume surrounded by the hypersphere, the target sample points are surrounded by the hypersphere as much as possible, while the
non-target sample points are excluded from the hypersphere as possible. Finally, this method achieves the purpose of dividing the two categories.

The optimization goal of SVDD is to find a sphere with a center and R radius. The optimization goal of SVDD is to find a minimum sphere whose center is a and whose radius is R.

\[ F(R, a, \xi) = R^2 + C \sum_i \xi_i \]  \hspace{1cm} (1)

Make this spherical surface satisfy:

\[ (x_i - a)^T (x_i - a) \leq R^2 + \xi_i \quad \forall_i, \xi_i \geq 0 \] \hspace{1cm} (2)

Where, \( \xi_i \) is a slack variable to prevent the model from over-fitting. C is a regularization factor that is used to adjust the magnitude of the effect of the relaxation variable.

3.2. Evaluation method of anomaly detection algorithm

The anomaly detection algorithm based on one-class learning is essentially a special two-class classification problem with only one type of data samples. For operation and maintenance data, normal samples do not need to be eliminated, and abnormal samples need to be eliminated. Suppose we have only two categories of classification targets, which are counted as positive and negative.

1) True positives (TP): the number of positive examples that are correctly divided into positive examples, that is, the number of instances (samples) that are actually positive examples and are classified into positive examples by the classifier;

2) False positives (FP): the number of positive examples that are incorrectly divided into positive examples, that is, the number of instances that are actually negative but are classified as positive by the classifier;

3) False negatives (FN): the number of instances that are incorrectly divided into negative examples, that is, the number of instances that are actually positive but are classified as negative by the classifier

4) True negatives (TN): The number of instances that are correctly divided into negative examples, that is, the number of instances that are actually negative and are classified as negative by the classifier.

The possible results of the classification are shown in Table I.

\[ \text{Tab. 1} \] Summary of possible results of the classification

| Actual category | Forecast category |  |
|-----------------|-------------------|---|
| Positive        | **True Positives(TP)** | **False Negatives(FN)** |
| Negative        | **False Positives(FP)** | **True Negatives(TN)** |

Different from the general classification algorithm evaluation index, here we mainly study two indicators: false alarm rate and false alarm rate. The so-called false alarm rate is the proportion of normal samples misjudged as abnormal FN/ (FN+TN). The rate of false alarm is FP/ (FP+TN) in which abnormal samples are misjudged as normal.TPR and FPR are classified correctly as the proportion of normal classes (TPR, True Positive Rate) from the point of view of actual classes.TPR = TP / (TP + FN) evaluates the correct proportion of normal samples classified by the algorithm, and the proportion of false classes classified as normal classes (FPR, False Positive Rate). FPR = FPR / (FP + TN) evaluates the proportion of abnormal samples classified into normal classes.

A good classifier should have a high detection rate, a low false alarm rate and a low false alarm rate. The ROC provides a dynamic performance observation of anomaly detection. By estimating the threshold of the anomaly detection decision variable, the dynamic game situation of the TPR and FPR indicators is estimated. AUC is an indicator that further measures the ROC performance of a one-class
classifier, which measures the area under the ROC curve and can be approximately calculated by the Wilcoxon-Mann-Whitney statistic:

$$ AUC = \frac{\sum_{i}^{n_{+}} \sum_{j}^{n_{-}} I(f(x_{i}^{+}) < f(x_{j}^{-}))}{n_{+}n_{-}} $$

(3)

Where, $x_{i}^{+}$ and $x_{j}^{-}$ represents a test sample whose true category is positive (negative), $n_{+}$ and $n_{-}$ is the number of true positive sample (negative sample) in test sample, $f(x_{i}^{+})$ and $f(x_{j}^{-})$ respectively represent the squared distance of c and d from the center of the hypersphere. I() is the indication function (the value is 1 when the condition is true, otherwise 0). And Ling [32] et al. prove that the AUC metric is a metric that is more discriminative than the accuracy metric, and thus it has become the main indicator for evaluating the performance of anomaly detection.

4. Experiment

4.1. Data set
This paper assumes that the business system is in stable operation. The collected data can be any IT asset running status, for example, the total CPU load rate. There are 144 collection points every day, collected every 10 minutes. The data collection period is 30 days, and the working days and non-working days are not distinguished. The data samples obtained are shown in Table II. Ten-fold cross-validation is used, that is, 90% of the data samples are used as training samples, and the remaining 10% are used as test samples.

| Tab. 2 Data sample description | |
|--------------------------------|---|
| the total CPU load rate | The number of samples |
| Positive | 1440 |
| Negative | 125 |

4.2. Experimental results and analysis
In this paper, the chi-square test algorithm is used as the feature selection model, and the KNN algorithm [17] is used as the comparison algorithm.

| Tab. 3 Abnormal Detection Rate of Total CPU Load Rate | |
|--------------------------------|---|
| Method | Abnormal detection rate (%) |
| SVDD | 95 |
| kNN | 81 |

As can be seen from Table III, the performance of the SVDD algorithm is better than KNN.
In the ROC curve, the normal detection rate TPR is usually used as the ordinate, and the false alarm rate FPR (FPR=1-TNR, that is, the missed alarm rate=1-anomaly detection rate) is the abscissa. By tracking the change in the threshold, a dynamic curve of the detection rate and the missed alarm rate of the normal class can be obtained. The detection rate of the normal class becomes low, which means that the unknown sample is not easily judged as a normal sample, but is more likely to be judged as an abnormal sample, so that the ratio of the abnormal class sample determined to be abnormal, that is, the abnormality detection rate becomes high. The rate of missed alarms becomes lower; the detection rate of normal classes becomes higher, which means that the algorithm tends to accept more samples as normal classes, so that the ratio of abnormal samples is determined to be abnormal, that is, the rate of abnormality detection becomes Low, the missed alarm rate becomes higher. Since the ideal detector
should reject all exception classes (FPR = 0) while accepting all normal classes (TPR = 1), the closer the ROC curve is to the upper left corner detector, the better the performance.

![Fig. 1 ROC curve of abnormal CPU total load rate](image)

**Fig. 1** ROC curve of abnormal CPU total load rate

It can be seen from Figure 1 that the ROC curve of the SVDD algorithm is closer to the upper left corner, indicating that its performance is better than the kNN algorithm with the ROC curve close to the bottom.

5. Conclusion

This paper assumes that the operation and maintenance data collected during the steady state of the system is mostly normal data. However, due to frequent changes in system services, the amount of data in a certain state is insufficient. This causes problems such as insufficient normal sample data, scarcity of faulty sample data, and lack of prior knowledge. Aiming at the above situation, an anomaly detection method based on one-class learning is proposed, and the classifier is trained by using limited normal data samples to guide the elimination of abnormal data. For the one-class classification problem, it is roughly divided into three categories: density estimation method, reconstruction method, and support boundary method. In this paper, the SVDD (Support Vector Data Description) method is used to eliminate the abnormal data in the collected operation and maintenance data, so that the subsequent data analysis can be better carried out. Experiments show that the proposed method is reasonable and effective.

**Acknowledgments**

This work was financially supported by the State Grid Corporation of Science and Technology (WBS number: 522723180006).

**References**

[1] Tax D M J. One-class classification: concept-learning in the absence of counter-examples [D]. Delft: Delft University of Technology 2001.

[2] Kosinski A S. A procedure for the detection of multivariate outliers [J]. Computational statistics & data analysis, 1998, 2: 145-161.

[3] Lauer M. A Mixture Approach to Novelty Detection Using Training Data with Outliers [A]. In Proceedings of the 12th European Conference on Machine Learning [C], 2001: 300-311.

[4] Yeung D-Y, Chow C. Parzen-window network intrusion detectors [A]. In Proceedings of the Sixteenth International Conference on Pattern Recognition [C], 2002: 385-388.

[5] Knorr E M, Ng R T, Tucakov V. Distance-based outliers: algorithms and applications [J]. The VLDB Journal, 2000, 8(3-4): 237-253.

[6] Jiang M F, Tseng S S, Su C M. Two-phasee clustering process for outliers detection [J]. Pattern Recognition Letters, 2001, 22(6-7): 691-700.
[7] Shyu M, Chen S, Sarinnapakorn K, etc. A novel anomaly detection scheme based on principal component classifier [A]. In Proceedings of the IEEE Foundations and New Directions of Data Mining Workshop [C], 2003: 12-21.

[8] Hoffmann H. Kernel PCA for novelty detection [J]. Pattern Recognition, 2007, 40(3): 863-874.

[9] Byungho H, Sungzoon C. Characteristics of auto-associative MLP as a novelty detector [A]. In International Joint Conference on Neural Networks, 1999 [C], 1999: 3086-3091.

[10] Thompson B B, Marks II R J, Choi J J, etc. Implicit learning in autoencoder novelty assessment [A]. In 2002 International Joint Conference on Neural Networks [C], 2002: 2878-2883.

[11] Owens J, Hunter A. Application of the self-organising map to trajectory classification [A]. In Proceedings of the Third IEEE International Workshop on Visual Surveillance [C], 2000: 77-83.

[12] Wong M L D, Jack L B, Nandi A K. Modified self-organising map for automated novelty detection applied to vibration signal monitoring [J]. Mechanical Systems and Signal Processing, 2006, 20(3): 593-610.

[13] Vapnik V N. The Nature of Statistical Learning Theory [M]. Berlin: Springer-Verlag, 1999.

[14] Schölkopf B, Platt J C, Shawe-Taylor J, etc. Estimating the Support of a High-Dimensional Distribution [J]. Neural Computation, 2001, 13(7): 1443-1471.

[15] Tax D M J, Duin R P W. Support Vector Data Description [J]. Machine Learning, 2004, 54(1): 45-66.

[16] TAX, David MJ; DUIN, Robert PW. Support vector data description. Machine learning, 2004, 54.1: 45-66.

[17] Keller J M, Gray M R, Givens J A. A fuzzy K-nearest neighbor algorithm [J]. IEEE Transactions on Systems Man & Cybernetics, 2012, SMC-15(4):580-585.