Research on Abnormal Detection Based on Improved Combination of K - means and SVDD

Hao Xiaohong¹, Zhang Xiaofeng²

¹School of Electrical Engineering and Information Engineering, Lanzhou University of Technology, Lanzhou 730050, China
²School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, China

Abstract. In order to improve the efficiency of network intrusion detection and reduce the false alarm rate, this paper proposes an anomaly detection algorithm based on improved K-means and SVDD. The algorithm first uses the improved K-means algorithm to cluster the training samples of each class, so that each class is independent and compact in class; Then, according to the training samples, the SVDD algorithm is used to construct the minimum superspheres. The subordinate relationship of the samples is determined by calculating the distance of the minimum superspheres constructed by SVDD. If the test sample is less than the center of the hypersphere, the test sample belongs to this class, otherwise it does not belong to this class, after several comparisons, the final test of the effective detection of the test sample. In this paper, we use KDD CUP99 data set to simulate the proposed anomaly detection algorithm. The results show that the algorithm has high detection rate and low false alarm rate, which is an effective network security protection method.

1. Introduction
With the rapid development of computer technology and Internet technology, intrusion has become a major threat to network security. In order to deal with network intrusion, network security experts in the field of a number of technical means, such as: firewall, data encryption and secure routing [1]. However, relying solely on a single static security technology such as firewalls can not resist complex and complex network attacks. As a result, experts in the field of network security proposed a network intrusion detection technology.

Intrusion detection technology is divided into two categories: misuse detection and anomaly detection [2]. Misuse detection refers to the use of specific types of intrusion behavior to create a feature library, when detected with the characteristics of the matching behavior is considered intrusion behavior [3]. But the method has a high false negative rate. Exception detection refers to the use of normal user behavior learning modeling, get the normal situation of the probability model [4]. Detection, the user behavior and the characteristics of the library to compare, the greater the deviation is regarded as intrusion. Based on the statistical anomaly detection is the traditional intrusion detection technology, it can only small-scale network detection, the face of large-scale network of its detection efficiency is low. Data mining technology can analyze the Internet data, from which to tap the useful information, and use this information to detect abnormal intrusion and known intrusion has become the most important network intrusion detection tool [5].

Wenke Lee et al [6] applied the data mining technology to the intrusion detection system for the first time. Through the use of data mining method to analyze a large number of audit data, through this method can get more accurate results, and its learning efficiency will be improved, can effectively
improve Intrusion detection system accuracy and expand its scope of application, but the use of this method requires a larger storage space. In order to solve this problem, Keerthi et al [7] proposed a decomposition algorithm that transforms the quadratic programming problem into a series of small-scale quadratic programming problems, which improves the learning rate and reduces the memory space requirements without affecting the classification accuracy, and A SMO algorithm is proposed, which can solve the optimal solution of each sub-problem and accelerate the convergence speed of the algorithm. However, this method increases the computational complexity. In order to further improve the learning speed of the algorithm, Collobert et al [8]. Proposed a parallel learning algorithm for sample set segmentation, which divides the sample set into P sub-samples by using the principle of divide and conquer, so as to improve the detection efficiency of the algorithm. However, since the P subsets are randomly divided, the accuracy of the classification is reduced, and the addition of additional rules increases the complexity of the algorithm. However, since the P subsets are randomly divided, the accuracy of the classification is reduced, and the addition of additional rules increases the complexity of the algorithm. In order to reduce the computational complexity, Cataltepe Z et al [9]. Proposed a semi-supervised decision tree algorithm for online feature selection, which uses online clustering to summarize the available network data, using extended cluster features to represent clusters. The characteristics of the original feature, but also the characteristics of the relationship between the description of the cluster, each cluster is marked as abnormal or normal by the user, and then based on the information to train the decision tree, according to the output of the decision tree to mark the input of new data. Although the computational complexity of the algorithm is reduced, the detection rate is still not high. In this paper, an improved anomaly detection algorithm combining K-means and SVDD is proposed. The basic idea of this algorithm is to cluster each category by K-means clustering algorithm first, and then build multiple hyperspheres from SVDD according to the result of clustering. By calculating the sample data to the minimum superspheres constructed by SVDD Distance to determine the affiliation of the sample. Finally, this paper uses the KDD CUP99 data set to simulate the proposed detection algorithm.

2. K-means algorithm and its improvement

2.1 K-means algorithm

K-means clustering algorithm is a typical partitioning method. It divides n objects into K clusters with K as input parameters, so that the clusters are independent and similar, and the similarity of clusters is The average of the objects in the cluster is calculated [10]. Although the computational complexity of K-means algorithm is low and easy to implement, there is a major drawback of this algorithm, which is sensitive to outliers [11]. Because a large extreme value of the object may significantly distort the data distribution, coupled with the K-means algorithm is to use the square error function to determine the convergence, making its sensitivity to the outlier increase, resulting in clustering results Inaccurate.

2.2 Improved K-means algorithm

In order to reduce the sensitivity of traditional K-means algorithm to outliers, an improved K-means algorithm is adopted in this paper. The basic idea of this algorithm is to represent the cluster by selecting the actual object in the cluster, and the remaining objects are allocated to the cluster with the most similar representative objects, so that the partitioning method can still be used to minimize all objects Corresponding to the degree of dissimilarity between the reference points [12]. At this point, the criterion function uses the absolute error criterion, which is defined as follows:

\[ E = \sum_{i=1}^{k} \sum_{j \in C_i} \| p - o_i \|^2 \]

Where \( E \) is the square error of all the objects in the data set, and the point in the \( p \) space, representing a given object in the cluster \( C_i \), and \( o_i \) is the representative object in the cluster \( C_i \).

The improved K-means algorithm has been iterated several times until each representative object becomes the actual center of its cluster. In the iterative process, as long as the clustering results continue to change, this iterative process will continue to use non-representative objects to replace the
representative object. In order to determine whether the non-representative object \( o_{non} \) is the best representative of the current representative, the following four cases need to be considered for each non-object representative:

1) \( p \) currently belongs to the representative object \( o_i \), and redistribute \( p \) to \( o_j \) if \( o_j \) is replaced by \( o_{random} \) as the new delegate object and \( p \) is closest to the other representing object \( o_i (i \neq j) \).

2) \( p \) currently belongs to the representative object \( o_i \), if the \( o_j \) is replaced by \( o_{random} \) as a new representative object, and \( p \) distance \( o_{random} \) is nearest, then \( p \) is reassigned to \( o_{random} \).

3) \( p \) currently belongs to the representative object \( o_i \), \( i \neq j \). If \( o_j \) is replaced by \( o_{random} \) as a new representative object, and \( p \) is still closest to \( o_j \), the membership of the object does not change.

4) \( p \) currently belongs to the representative object \( o_i \), \( i \neq j \). If \( o_j \) is replaced by \( o_{random} \) as a new representative object, and \( p \) is closest to \( o_{random} \), then \( p \) is reassigned to \( o_{random} \).

The difference in absolute error has an effect on the cost function when each reallocation occurs. Thus, if the current representative object is replaced by a non-representative object, the cost function calculates the difference in absolute error. The total cost of the exchange is the sum of the costs of all non-representative objects. When the total cost is negative, the actual absolute error \( E \) will be reduced, \( o_j \) can be replaced by \( o_{random} \); when the total cost is positive, then the current representative of the object \( o_j \) is reasonable, in this iterative process will not change.

The improved K-means algorithm steps are as follows:

1) arbitrarily select \( K \) objects from \( D \) as the initial representative;
2) Repeat
3) assign each remaining object to the cluster represented by the nearest delegate object;
4) arbitrarily select a non-representative object \( o_{random} \);
5) calculate the total cost of representing the object \( o_j \) with the \( o_{random} \) switch;
6) If \( S < 0 \), then \( o_j \) is replaced by \( o_{random} \) to form a new set of \( K \) representative objects;
7) until the cluster no longer changes.

3. An Algorithm for Abnormal Detection Based on Improved K - means and SVDD

The anomalous detection algorithm based on improved K-means and SVDD proposed in this paper is based on two considerations [13]: First, in most cases, the normal data set is far redundant with abnormal data sets; Second, the imbalance between normal and abnormal data makes it possible to train the model when the training results may lead to deviation, affecting the test results. In this paper, we first use the improved K-means algorithm to cluster the training samples of each class so that each cluster is similar to each other, and then the result of clustering is used to train the SVDD algorithm. Clustering support vector, and finally all the individual support vector combination retraining to get the global optimal solution. Support vector domain description (SVDD) algorithm is a kind of support vector machine (One-Class SVM) proposed by Tax [14] and so on in the traditional support vector machine. The idea of the algorithm is to use the sample to be classified as a whole, to establish a closed area, so that all the samples belonging to the class all or as much as possible in this closed area, and do not belong to the class of exclusion In this closed area, the sample on the region is the support vector obtained by SVDD.

When a normal data set \( \{x_1, x_2, ..., x_n\} \) is given, which contains \( n \) data objects, as a training sample for constructing the One-class classifier, an attempt is made to find a minimum volume of the hypersphere (center \( \alpha \), radius \( K \)) so that all or as much as possible with \( x_i \) is contained within the supersphere [15]. In general, the training samples are more or less containing noise, so the above optimization results are more sensitive to these noise, lack of robustness [16]. In order to improve the robustness of the algorithm, the relaxation variable \( \xi_i \geq 0 \) is introduced for each sample \( x_i \). Minimizing the volume of the
hypersphere is the quadratic programming problem, that is minimizing \( \min \left[ R^2 + C \sum \xi_i \right] \) under the constraint condition \( (x_i - a)^2 \leq R^2 + \xi_i \).

Where \( C \) is a specified constant, its role is to control the degree of punishment on the wrong sample, to achieve the wrong proportion of the sample and the complexity of the algorithm between the compromise. Using the Lagrangian function to solve the minimum problem under the above constraints, we can get the dual problem of the original problem:

\[
L = \sum a_i K(x_i, x_j) - \sum a_i a_j K(x_i, x_j)
\]

In the above formula, the \( a_i \) should conform to the constraints: \( \sum a_i = 1 \) and \( 0 \leq a_i \leq C \), and \( K(x, x) \) is the inner product of the above formula, so \( K(x, x) \) should satisfy the kernel function of Mercer condition \([17]\).

By the KKT condition, the sample corresponding to the \( a_i = 0 \) is in the hypersphere; the corresponding \( 0 \leq a_i \leq C \) sample is called the support vector on the hypersphere, which is regarded as an anomaly point with the partial sample corresponding to the \( a_i > C \).

When a new sample \( Z \) is given, it is judged whether the sample belongs to the target sample, and the corresponding discriminant function is:

\[
\| z - a \|^2 = K(z, z) - 2 \sum a_i K(x_i, z) + \sum a_i a_j K(x_i, x_j) \leq R^2
\]

If the above discriminant function holds, then \( Z \) belongs to the target sample and belongs to that class; otherwise \( Z \) is not the target sample and does not belong to that class.

4. Simulation experiment

4.1 Experimental data set

In this paper, we use KDD CUP99 \([18]\) data set to detect the proposed intrusion detection method. The data set includes four types of attack types: DoS, Probe, R2L, and U2R. In the KDD CUP99 data set, each network connection consists of 41 feature attributes and 1 flag attribute, of which there are 7 discrete feature attributes and 34 consecutive feature attributes \([19]\). In order to make the test results are universal, this paper uses three sets of data sets to carry out experiments, experimental data shown in Table 1.

| Data set | Normal | DoS | Probe | R2L | U2R |
|----------|--------|-----|-------|-----|-----|
| 1        | 500    | 400 | 350   | 200 | 40  |
| 2        | 450    | 300 | 300   | 150 | 30  |
| 3        | 400    | 200 | 280   | 100 | 25  |

4.2 Experimental results and analysis

In this paper, the proposed algorithm is tested by simulation experiment. The experimental results are shown in Table 2.

| Attack | K-means | SVDD | Proposed |
|--------|---------|------|----------|

Table 1 Experimental data

Table 2 Comparison of experimental results
| ck type  | Detection rate (%) | False alarm rate (%) | Detection rate (%) | False alarm rate (%) |
|---------|--------------------|----------------------|--------------------|----------------------|
| DoS     | 91.25              | 6.37                 | 93.9               | 7                   |
|         |                    |                      |                    | 98.64               | 4.52                |
| Probe   | 93.74              | 5.78                 | 96.2               | 1                   |
|         |                    |                      |                    | 97.81               | 3.85                |
| R2L     | 73.51              | 25.2                 | 84.6               | 2                   |
|         |                    |                      |                    | 89.42               | 20.2                |
| U2R     | 90.38              | 8.94                 | 94.3               | 5                   |
|         |                    |                      |                    | 95.54               | 6.97                |

It can be seen from Table 2 that the improved K-means and SVDD proposed intrusion detection algorithm compared with the other two methods of DoS, Probe, U2R and other three types of attack has a high detection rate, but R2L this type of attack detection rate is low, false alarm rate of 20.21%. This is because the sample of this attack is less than other types of attacks, and many of the attacks in R2L have high similarity to normal data, so that some attacks are included in the normal type, resulting in a relatively low detection rate.

![Figure 1. Comparison of detection rates for different algorithms](image)

Figure 1 shows that the proposed intrusion detection algorithm combined with the improved K-means and SVDD is compared with the commonly used intrusion detection algorithm. It can be seen that the proposed method has obvious advantages over other detection methods in detecting detection rate. The algorithm based on BP neural network and the algorithm proposed in this paper are relatively close to the detection rate, but the former computational complexity is relatively high, so it takes a long training time.

5. Conclusion
In this paper, an improved K-means and SVDD algorithm is proposed to solve the problem of low detection rate and high false alarm rate in network intrusion detection. Firstly, the algorithm is clustered by the improved K-means clustering algorithm. Then, a number of superspheres are established by SVDD according to the result of clustering. By calculating the distance between the minimum superspheres constructed by SVDD Determine the affiliation of the sample, if the test sample within the hypersphere, the sample belongs to the class, otherwise it does not belong to the class. The experimental results show that the proposed algorithm has high detection efficiency and low false alarm rate for different attacks, which has obvious advantages over traditional intrusion detection methods. However, the intrusion detection algorithm proposed in this paper also has some
shortcomings, such as when the attack type of sample data is less detection efficiency is not too ideal. Therefore, in the future research work, the use of a variety of classification algorithms to build intrusion detection system will be a hot research.

6. References

[1] Chetan R, Ashoka D V. Data mining based network intrusion detection system: A database centric approach [C]//Computer Communication and Informatics (ICCCI), 2012 International Conference on. IEEE, 2012: 1-6.

[2] Pan Z S, Chen S C, Hu G B, et al. Hybrid neural network and C4. 5 for misuse detection[C]//Machine Learning and Cybernetics, 2003 International Conference on. IEEE, 2003, 4: 2463-2467.

[3] Guo C, Ping Y, Liu N, et al. A two-level hybrid approach for intrusion detection[J]. Neurocomputing, 2016, 214: 391-400.

[4] Kim G, Lee S, Kim S. A novel hybrid intrusion detection method integrating anomaly detection with misuse detection [J]. Expert Systems with Applications, 2014, 41(4): 1690-1700.

[5] Jaiganesh V, Mangayarkarasi S, Sumathi P. Intrusion detection systems: A survey and analysis of classification techniques [J]. International Journal of Advanced Research in Computer and Communication Engineering, 2013, 4(2): 1629-1700.

[6] Lee W, Stolfo S J. Data Mining Approaches for Intrusion Detection[C]//Usenix security. 1998.

[7] Keerthi S S, Gilbert E G. Convergence of a generalized SMO algorithm for SVM classifier design [J]. Machine Learning, 2002, 46(1-3): 351-360.

[8] C. F. R., B. S. M., P. T. S. A., J. 0. A. D. Finding a mixture of large SVEs for very large scale problems[J]. Neural computation, 2002, 14(5): 1105-1114.

[9] Cataltepe Z, Ekmeki U, Cataltepe T, et al. Online feature selected semi-supervised decision trees for network intrusion detection[C]//Network Operations and Management Symposium (NOMS), 2016 IEEE/IFIP. IEEE, 2016: 1085-1088.

[10] Muniyandi A P, Rajeswari R, Rajaram R. Network anomaly detection by cascading k-Means clustering and C4. 5 decision tree algorithm [J]. Procedia Engineering, 2012, 30: 174-182.

[11] Patel J, Panchal K. Effective Intrusion Detection System using Data Mining Technique [J]. Journal of Emerging Technologies and Innovative Research, 2015, 2 (6).

[12] Al-Yaseen W L, Othman Z A, Nazri M Z A. Multi-level hybrid support vector machine and extreme learning machine based on modified K-means for intrusion detection system [J]. Expert Systems with Applications, 2017, 67: 296-303.

[13] Huang Tao, Liu Shenghui, Tan Yanna. Based on k-means clustering algorithm [J]. Computer Technology and Development, 2011, (07): 54-57 + 62.

[14] Tax D M J, Duin R P W. Support vector domain description [J]. Pattern recognition letters, 1999, 20(11): 1191-1199.

[15] Huang Qian, Wang Zhen, Wei Tao, Chen Yu. Real-time intrusion detection system based on One-class SVM [J]. Computer Engineering, 2006, (16): 127-129.

[16] Perdisci R, Gu G, Lee W. Using an ensemble of one-class svm classifiers to harden payload-based anomaly detection systems[C]//Data Mining, 2006. ICDM’06. Sixth International Conference on. IEEE, 2006: 488-498.

[17] Fisher W D, Camp T K, Krzhizhanovskaya V V. Anomaly detection in earth dam and levee passive seismic data using support vector machines and automatic feature selection [J]. Journal of Computational Science, 2017, 20: 143-153.

[18] Kosher J, Bag M. Cascading of C4. 5 decision tree and support vector machine for rule based intrusion detection system [J]. International Journal of Computer Network and Information Security, 2012, 4(8): 8.

[19] Muniyandi A P, Rajeswari R, Rajaram R. Network anomaly detection by cascading k-Means clustering and C4.5 decision tree algorithm [J]. Procedia Engineering, 2012, 30: 174-182.