Intrusion Detection Method Based on Improved K-Means Algorithm

Chun Meng¹, Yuanyuan Lv¹*, Long You² and Yuchen Yue¹

¹School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China
²Academy of Intelligent Collaborative Cloud System, China Aerospace Science and Industry Corporation Limited, Beijing 100048, China
*Corresponding author: 694746485@qq.com

Abstract. Data mining technology has a good application in the field of intrusion detection. For the problem that K-Means algorithm is difficult to process high-dimensional data, local optimal solution, and cannot determine K value, this paper proposes an improved K-Means algorithm. Firstly, the PCA algorithm is used to reduce the dimension of the data set, and then the Outlier detection is used to eliminate the Outliers that have a great influence on the final clustering result. Then, the initial clustering center point is selected based on the distance to avoid the local optimal solution. Finally, K-means algorithm performs clustering to obtain an intrusion cluster. The experimental results show that compared with the common data mining-based intrusion detection algorithm, the proposed method has a good performance in detection rate and false positive rate, and its performance has also improved.

1. Introduction

In recent years, with the rapid development of the Internet, a large number of network security incidents such as personal information theft, Trojan virus, Botnet, flood attack, and network malicious programs appear frequently. Network security requires a more effective dynamic defense technology [1]. At the same time, data mining algorithms are continually optimized and improved to meet the needs of dynamic intrusion detection technology. Data mining algorithms can help analyze historical data, find patterns in seemingly chaotic data, make it evidence-based, and provide a more accurate detection basis for intrusion detection [2].

At present, computer science research on data mining-based intrusion detection technology is relatively mature, but there are still many unsolved problems. First of all, algorithm selection, K-means, svm, decision tree, neural network and many other mature data mining algorithms can be used to construct intrusion detection models [3], but they all have obvious defects, which need further exploration and improvement. Therefore, the requirements for the selection and improvement of the algorithm are very high when establishing the behavioural contour. Secondly, in terms of feature selection, there is currently no recognized algorithm with very high efficiency to help data mining algorithms to reduce data. Based on this, this paper proposes an improved algorithm based on machine learning clustering algorithm K-means, and also features feature selection and dimension reduction methods and correlation analysis algorithms, and applies it to intrusion detection.
2. Intrusion Detection based on Improved K-Means Algorithm

2.1. The Deficiency of K-Means Algorithm Applied to Intrusion detection
K-Means is a classical algorithm in clustering algorithm [4], which has a good application effect in intrusion detection, but there are some shortcomings in this algorithm.

First of all, as the data dimension increases, the overhead of the K-Means algorithm will increase rapidly, resulting in reduced algorithm efficiency.

Secondly, the K-Means algorithm is very sensitive to data noise and discrete points. Because it will iterate the cluster center point obtained in the previous cycle calculation into the next round operation, and recalculate the new cluster center point. However, if there are some distant noise points in the data set, it will be greatly affected when recalculating the sum of error squares and cluster centers. These erroneous cluster centers will increasingly deviate from the really dense data sets in the clusters and slowly move closer to the farther noise points, which can lead to inaccurate results. Therefore, before the K-Means algorithm is executed, the noise points in the data set should be eliminated first, and the clustering accuracy of the K-Means algorithm should be improved.

Thirdly, the initial clustering center of the algorithm is randomly selected, and the selection of the algorithm and the clustering center is strongly correlated. The wrong clustering center selection may cause the algorithm to fall into the local optimal solution, resulting in the final clustering failure. The K-Means algorithm is a relocation technique that continually iterates until the optimal solution is found. The initial clustering center is randomly selected, thus causing the instability of the algorithm results. In addition, the K-Means algorithm uses the sum of squared errors, that is, the sum of the Euclidean distances from all points in the cluster to the cluster center, as a criterion function to measure the degree of cluster completion. But mathematically, the sum of squared errors belongs to a non-convex function, that is, there may be multiple minimum values with a reciprocal of zero in the function, but the minimum value of the whole function is only one. If the initial clustering center of the K-Means algorithm falls within the local convex function interval, or falls within a non-convex function surface, it may lead to the optimal solution search error, and only finds a function extreme point rather than the entire function. The minimum point has a very large impact on the final result.

2.2. Improvement of K-Means Clustering Algorithm

2.2.1 Feature extraction based on principal component analysis. Due to the complex network environment, the data characteristics of the network access acquired by the intrusion detection system tend to be higher in dimension, that is, the data features in the data set are too many. These high-dimensional data features often contain more redundant features that do not help to determine whether the unknown behavior is normal or intrusive. More critically, the performance of the K-Means algorithm is very susceptible to high-dimensional data, forming a dimensional disaster. High-dimensional data seriously drags down the efficiency of the algorithm. Therefore, it is necessary to reduce the dimension of the original data set, which is a key step to improve the efficiency of intrusion detection. To this end, this paper chooses the principal component analysis algorithm (PCA) to help data mining algorithm for feature dimension reduction [5].

When performing the PCA algorithm and the clustering algorithm, the input features must be numeric data. However, the intrusion detection data set represented by KDD CUP99 contains data such as strings and Boolean data that do not meet the input requirements. Therefore, the data set needs to be normalized before executing the algorithm. The specific method is as follows:

(1)Continuous data is legal data, but if the range of values of these features is too large or too small, or the units of the features are different, it will cause problems in the clustering results. To this end, we have to fix the characteristics of the continuous range of values and normalize them. In this paper, the Min-Max normalization (linear normalization) method is adopted. The principle is that the data is linearly processed, and the difference between the maximum and minimum values of the feature is used as the denominator reference, and the difference between the eigenvalue and the feature minimum. The value is taken as a numerator and its ratio is used as the new eigenvalue. In this way,
features that are too large or too small in the original range can be mapped into a closed interval of 0 to 1. The function is expressed as follows:

\[ x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \]  

(1)

If the feature value of the newly added sample is not in the existing maximum and minimum values of \( x \), the feature values greater than the existing maximum value are all classified as 1, and the feature values smaller than the existing maximum value are classified as 0. For continuous data, the distance between two points is calculated using the traditional Euclidean distance. The formula is as follows:

\[ d(x_i, x_j) = \left( \sum_{k=1}^{d} (x_{ik} - x_{jk})^2 \right)^{\frac{1}{2}} \]  

(2)

For discrete variables, the characteristic range of data is several specific values, including character discrete type and digital discrete type. Since the PCA algorithm and the K-Means algorithm can only recognize numeric data, it is necessary to map character discrete data to numbers. A Boolean value is also a type of discrete data that is directly converted to 0 and 1 for this two separated data. If the data is multi-separated type data, for example, some discrete data has four values of A, B, C, and D, then the interval of 0 to 1 is equally divided into four, and the value of A is 0, the value of B is 0.33, the value of C is 0.66, and the value of D is 1. N separates the bulk data and so on.

Unlike continuity data, the distance of discrete data uses a simple matching coefficient as the distance formula for discrete data. Taking the two separated scattered data as an example, the formula is as follows:

\[ SMC = (b + c)(a + b + c + d)^{-1} \]  

(3)

In the entire sample set, \( a \) represents the number of times the two samples have this attribute at the same time. \( b \) indicates the first attribute of the two samples is 1, and the second data has the number of times this attribute is 0. \( C \) indicates the number of times the first attribute of this attribute is 0 and the second data attribute is taken as 1 for two samples. \( d \) indicates the number of times the two samples are 0 at the same time.

(3) For the mixed data, using the two methods mentioned above, the continuous data is normalized by the linear normalization method, and the discrete data is normalized by the method mentioned in (2). This way all data is mapped to a value range of 0 to 1. Then calculate the distance between the two sample sets, the formula is as follows:

\[ d(x_i, x_j) = \sum_{k=1}^{d} d_{ij}^{(k)} \]  

(4)

\[ d_{ij}^{(k)} \begin{cases} |x_{ik} - x_{jk}|, & \text{Continuous type} \\ 1, & \text{Discrete type and } x_{ik} \neq x_{jk} \\ 0, & \text{Discrete type and } x_{ik} = x_{jk} \end{cases} \]  

(5)

The PCA algorithm is used to reduce the data set to an acceptable dimension range, improve the efficiency of the K-Means algorithm, and reduce the impact of redundant features on the final clustering result.

2.2.2. Outlier detection. In the foregoing discussion on the deficiencies of the K-Means algorithm, it can be seen that there may be some abnormal data in the data set with a large distance from the normal cluster, or an Outlier. These Outliers have a great impact on the final clustering results. These Outliers are of a smaller magnitude relative to normal behaviour in network behaviour. Therefore, it can be assumed that the Outliers detected by the algorithm belong to the abnormal data. This paper chooses a density-based Outlier detection algorithm. [6].
First of all, define some concepts in Outlier detection, as shown in Table 1.

| Name                        | Definition                                                                 | Regularity                                                                 |
|-----------------------------|----------------------------------------------------------------------------|----------------------------------------------------------------------------|
| k-distance                  | Generally represented by d(p,o), it refers to the radius length of a circle containing k data points closest to the data point p. | The larger the K distance, the smaller the data density, otherwise the higher the density |
| \(N_{kd}(p)\) K Distance   | The K-distance neighbourhood of data point p represents a set of all points in all data points that are less than or equal to K distance from p. | Outliers are positively correlated with K distance neighbourhood |
|\( \text{Reach-dist} k(p,o) \) Reachable distance | The greater between the distance between the two points p and o and the distance between the o and K points. Reach-dist \( k(p,o) = \max \{d(p,o), k\text{-distance}(o)\} \) | If the distance between p and o is greater than the distance between o and K, the reachable distance is K distance, otherwise the reachable distance is the distance between p and o. |
| \( \text{Lrdk}(p) \) Local reachable density | The local reachable density of P describes the degree of Outliers of P. It is the reciprocal of the average of the reachable distances of all data points in the K neighbourhood. | Local reachable density is proportional to the Outliers |
| \( \text{LOF}(p) \) Local Outlier | Used to describe the Outliers of point P. \( \text{LOF}_k(q) = \frac{\sum_{q \in N_k(q)} \text{lr}(p)}{|N_k(q)|} \) | It is proportional to the degree of Outliers. The greater the degree of Outliers, the larger the LOF, and vice versa. If the degree of Outliers is very low, the LOF value is infinitely close to 1. |

As can be seen from the above table, the LOF value is the core of the algorithm. The density-based Outlier detection algorithm rejects points where the density value is greater than the local Outlier by more than a given threshold, leaving points within the threshold to achieve the ultimate goal. The specific flow of the density-based Outlier detection algorithm is as follows:

Input: original data set, initial neighbour number num P, number of Outliers k
Output: K Outliers

1. Calculate the Euclidean distance between individual data points.
2. The Euclidean distances are sorted in descending order and the NO,K and K distance neighbourhoods are calculated.
3. Calculate the local reachable density.
4. Calculate the local Outliers.
5. The K Outliers with the largest K before the output are output according to the local Outlier factor.

The time complexity of LOF algorithm is O(n2), and the time complexity is acceptable compared with K-Means algorithm. Therefore, the density-based Outlier detection method is chosen as the method to eliminate the influence of Outliers on K-Means algorithm.

2.2.3. Improved K-Means clustering center selection. As can be seen from the foregoing discussion, the choice of the initial center point will affect the final clustering result. This paper proposes an initial clustering center point selection method and applies it to intrusion detection so that the K-Means clustering algorithm does not fall into the local optimal solution.

The principle of the K-Means algorithm is to cluster data with high similarity in the data set into one cluster, and the similarity between the data points of different clusters is low. Therefore, whether the distance between the cluster and the cluster center point is large enough is a key factor in judging the clustering result. However, if the distance between the center points is set large enough when the initial cluster center point is selected, it is easy to make the final clustering result avoid the local optimal solution in the iterative process. Therefore, the improved K-Means clustering algorithm selects the distance from the initial cluster center point.[7].
First of all, select the two points farthest from the Euclidean distance in all the data in the data set as two points in the K cluster center points. The remaining data in the data set is divided into clusters represented by the center points of the two clusters according to the distance of the Euclidean distance, that is, which cluster is into which the data set is closer to which initial center point. After completing the dividing step, count the number of points of the two clusters. Select the cluster with more data points and name it N. The cluster center corresponding to the N cluster is deleted, and the point with the farthest Euclidean distance is taken in N as the second and third initial cluster centers, and the above operation is repeated. This step is performed recursively until the NO.K initial cluster center point is found. The K initial cluster center points found at this time can be used as the input of the initial cluster center point of the K-Means algorithm, without having to randomly select K initial points by the algorithm itself. This method is based on the largest Euclidean distance between two points to find the initial center point, which can make these initial points relatively distant from each other, the similarity is low, and the possibility of local optimal solution is reduced.

Assuming the original data set is \( X = \{x_i \mid x_i \in \mathbb{R}^d, i=1, 2...n\} \), the data set contains n pieces of data, each of which has a feature dimension of m. The improved K-Means initial center point selection algorithm flow is as follows:

1. Enter the number K of the initial cluster center points.
2. Set K cluster clusters that can contain n data points, put all the data in the data set into one cluster, and set other clusters to be empty.
3. Calculate the number of data points in each cluster, and select the cluster with the most data points, and set it to Y.
4. Use Euclidean distance to calculate the distance of all points in Y, select the two data points with the farthest distance, and use these two points as the new cluster center point, and classify the cluster cluster with the number of cluster points still zero. Medium, set to A and B.
5. All points in Y are assigned to A and B according to the distance of the Euclidean distance.
6. It is checked whether all the K clusters have been allocated. If all the K clusters contain data points, the algorithm ends, and all K cluster centers are taken as the initial cluster center input of the K-Means algorithm. If there are still empty clusters, continue the recursive execution algorithm and jump to step 3.

2.3. Improved K-Means algorithm description
According to the above improvement of the K-Means algorithm, the flow of the entire improved K-Means algorithm is as shown in figure 1.
Figure 1. Improved K-Means algorithm flow.

Data preprocessing is performed on the collected data or the original data set. Data cleaning refers to the deletion or modification of data that has problems during the collection phase (including incomplete data, obvious errors in data, etc.). Data normalization uses the Min-Max linear function normalization to map data into intervals of 0 to 1. Feature extraction uses the PCA algorithm to perform feature dimension reduction on the entire data set.

Then the Outlier detection analysis of the whole data set will affect the removal of Outliers of K-Means cluster clusters and cluster center points, and improve the accuracy of clustering algorithm. Then using the center point selection method, the initial cluster center point is extracted and passed as an input to the K-Means algorithm. The K-Means algorithm ends after the iteration is stable, and outputs each cluster, which includes a normal behavior cluster and several abnormal intrusion behavior clusters.

3. Experimental Analysis of Application in Intrusion Detection

In this paper, the improved K-Means clustering algorithm is used in intrusion detection, and its application effect is verified by experiments.

3.1. Experimental Data Set

The data set used in the experiment is KDD CUP99 [8]. This article uses a 10% KDD CUP99 data set. The data set includes a test subset and a training subset extracted in the original KDD CUP99 data set. The training subset is used in the training phase of cluster analysis. The training subset is trained by the improved PCA-K-Means algorithm to obtain each cluster cluster and cluster center. The test subset is used to verify the experimental results of the algorithm and calculate the final detection rate and false positive rate. The data set includes two main types of intrusions, DOS and Probe, each of which has several specific attack types. In order to test the improved algorithm's ability to detect unknown intrusion types, the test subset also includes some attack modes that do not exist in the training subset. The unknown attacks that exist in the test subset are named Unknown.
3.2. Experimental Results and Analysis

The detection rate and false positive rate of various attack types in the test set are shown in Table 2:

| Intrusion type | Invasive subclass | Detection rate | False alarm rate |
|----------------|------------------|----------------|------------------|
| Normal         | Normal           | 99.02%         | 1.144%           |
|                | back             | 94.67%         | 4.000%           |
|                | land             | 93.75%         | 6.250%           |
|                | neptune          | 97.18%         | 2.817%           |
| DOS            | smurf            | 97.14%         | 2.857%           |
|                | pod              | 95.80%         | 3.497%           |
|                | teardrop         | 90.09%         | 3.846%           |
|                | ipsweep          | 94.49%         | 3.226%           |
|                | nmap             | 96.48%         | 2.232%           |
| Probe          | portsweep        | 94.49%         | 2.439%           |
|                | satan            | 97.87%         | 2.128%           |
| Unknown        | Unknown          | 92.59%         | 2.492%           |

As can be seen from the above table, the improved PCA-K-Means algorithm is excellent for the recognition of normal access behaviour. The detection rate reached 99.02%, and the false positive rate was only 1.144%. In terms of intrusion detection, the algorithm has good recognition ability for DOS and Probe attacks. The average detection rate for DOS intrusion reached 96.82%, and the average false positive rate was 3.761%. The average detection rate for Probe attacks was 95.47%, and the average false positive rate was only 2.733%. In addition, for the unknown intrusion behaviour, the improved algorithm also has a high recognition rate; the detection rate reaches 92.59%, and the false positive rate is 2.492%.

3.3. Comparative Experiment and Analysis

There are many common intrusion detection models based on data mining. Let's compare the experimental results of this algorithm with some common data mining algorithms. The most common is the original K-Means algorithm directly applied to intrusion detection, and the PSO-K mean algorithm combined with particle swarm algorithm and K-Means algorithm[9], PSO-SVM algorithm based on particle swarm optimization and SVM algorithm[10], NK-Means algorithm based on improved K-Means algorithm[11]. The algorithm described in this paper is named the PCA-NK mean algorithm. The comparison of the detection rates of the five algorithms for various intrusion types is shown in Table 3:

| Intrusion type | K-Means | NK-means | PSO-K mean value | PSO-SVM | PCA-NK mean value |
|----------------|---------|----------|------------------|---------|------------------|
| DOS            | 78.35   | 83.96    | 82.63            | 89.32   | 96.82            |
| Probe          | 80.15   | 87.32    | 89.36            | 92.54   | 95.47            |
It can be seen that compared with the other four algorithms, the proposed PCA plus improved K-Means algorithm has a higher detection rate for both DOS and Probe attacks.

The false alarm rates of the five algorithms are shown in Table 4:

| Intrusion type | K-Means | NK-means | PSO-K mean value | PSO-SVM | PCA-NK mean value |
|----------------|---------|----------|------------------|---------|------------------|
| DOS            | 8.135   | 4.693    | 4.983            | 3.998   | 3.761            |
| Probe          | 9.142   | 5.720    | 3.694            | 3.298   | 2.733            |

In the comparison of false positive rate, the performance of the PCA-NK mean algorithm is similar to that of the detection rate, with a lower false positive rate.

In terms of time complexity, the time complexity of the K-Means algorithm is $O(5nT)$. Where $T$ is the number of recursive algorithms and $n$ is the total number of samples in the clustered dataset. The principal component analysis algorithm used in this paper has a time complexity of $O(n)$, which is on the same order of magnitude as the K-Means algorithm. In the step of Outlier detection, since the distance between points needs to be calculated, the time complexity will reach $O(n^2)$. In the initial center point selection, the distance-based initial center point selection method also needs to calculate the distance of each point, so the time complexity is also $O(n^2)$. Therefore, in summary, the overall time complexity of the PCA-NK mean algorithm proposed in this paper is improved to $O(n^2)$ compared to the K-Means algorithm. The time taken by each algorithm after analyzing all the data is shown in Table 5:

| Algorithm   | K-Means | NK-Means | PSO-K mean value | PSO-SVM | PCA-NK mean value |
|-------------|---------|----------|------------------|---------|------------------|
| Time        | 273.3s  | 302.1s   | 251.7s           | 215.3s  | 211.9s           |

It can be seen that although the PCA feature dimension reduction, Outlier analysis and initial center point selection are added in the steps, the PCA-NK mean algorithm is still better than the original K-Means algorithm in execution time. Mainly because of the processing of the feature reduction in the PCA algorithm, the overall algorithm is reduced in time consumption. This shows that the PCA principal component analysis algorithm can not only improve the accuracy of the algorithm, but also improve the efficiency of the algorithm.

4. Conclusion
The essence of intrusion detection is to detect whether an unknown behaviour belongs to attack behaviour from a large number of network data packets according to known laws, which is very consistent with the idea of data mining. Therefore, a large number of intrusion detection systems now use data mining technology as a modeling method for their core detection models. In this paper, K-Means algorithm is selected in the process of intrusion detection modeling. For the problem of K-Means algorithm applied to intrusion detection, the K-Means algorithm is improved, and a PCA based improvement is proposed. K-Means method. The experimental results of the proposed clustering analysis algorithm show that the proposed algorithm is feasible, and it has a good effect on improving the detection rate of intrusion detection and reducing the false alarm rate.

5. Acknowledgement
This work was supported by the National Key Research & Development Program of China (2016QY06X1205).
6. References

[1] Viktor Mayer-Schonberger, Kenneth Cukier. Big data: a revolution that will transform how we live, work and think. Boston: Houghton Mifflin Harcourt, 2013.

[2] Romanovs. Building an Intrusion Detection System for IT Security Based on Data Mining Techniques [J]. Scientific Journal of Riga Technical University. Computer Sciences, 45(1), 2011.

[3] Wang Youjin. Application Research of Data Mining in Intrusion Detection System [D]. Jilin University, 2017.

[4] Zheng Z X, Wang M. The application of K-means clustering algorithm based on big data in network security detection [J]. Journal of Hubei University of Education, 2016.

[5] Ju F, Sun Y, Gao J, et al. Image Outlier detection and feature extraction via L1-norm based 2D probabilistic PCA [J]. IEEE Transactions on Image Processing, 24(12):4834-4846, 2015.

[6] Qin Hao. Density-based local outliers mining and its application in intrusion detection [D]. Dalian Maritime University, 2016.

[7] Feng Bo, Hao Wenning, Chen Gang, Zhan Donghui. Optimization of initial clustering center selection for K-means algorithm [J]. Computer Engineering and Applications, 49(14): 182-185, 2013.

[8] Zhang Xinyou, Zeng Huazhen, Jia Lei. Intrusion Detection Dataset KDD CUP99 Research [J]. Computer engineering and design, 31(22): 4809-4812, 2010.

[9] Del V Y, Venayagamoorthy G K. Particle swarm optimization: basic concepts, variants and applications in power systems [J]. IEEE Transactions on Evolutionary Computation, 12(2):171-195, 2008.

[10] Xiang Changsheng, Zhang Linfeng. Application of PSO-SVM in Network Intrusion Detection [J]. Computer engineering and design, 34(4): 1222-1225, 2013.

[11] Yan Yihan. Research on Intrusion Detection Method Based on k-means [D]. Harbin Institute of Technology, 2017.