The Improved K-means Algorithm in Intrusion Detection System

Research

Hongbo Zhang¹,a, Yi Jiang²,b

¹Department of Computer Science, Xiamen University, Xiamen, China
²Department of Computer Science, Xiamen University, Xiamen, China
a bohongzhang86@sina.com, b jiangyi@xmu.com

Keywords: Internet security, intrusion detection, k-means

Abstract. To improve the efficiency of Internet intrusion detection, data mining is adopted in intrusion detection. The paper introduces the concept of intrusion detection and k-means algorithm. For the defect of K-means algorithm, it proposes an improved K-means algorithm. Experiments show that the improved k-means algorithm can get a better detection rate.

Introduction

With the rapid development and widespread use of the Internet, while people benefit from the Internet, the Internet has also become the target of many malicious attacks. Internet intrusion detection is an important protection measure for Internet information security, which is able to detect unauthorized or unusual system behaviors and to alert the users’ attention to guard against. In this paper, the data mining method is applied to Internet intrusion detection to detect the intrusion, and provide real-time network security protection.

Intrusion Detection

Definition of Intrusion Detection. Intrusion detection is a process to identify an attempt to invade, an ongoing invasion or the invasion process has already taken place. It collects and analyzes information from key points of a computer network or system and responds if breaches of security policy and signs of attack are detected.

Types of Intrusion Detection. According to the test data source, intrusion detection system can be divided into host-based intrusion detection system and network-based intrusion detection system[2]. Host-based intrusion detection system is mainly concerned with detecting users’ behavior on the host. Network-based intrusion detection system is mainly about detecting network attacks.

According to the different detection angle, intrusion detection methods can be divided into anomaly detection and misuse detection[2]. Anomaly detection assumes the attacker's behaviors different from the normal behaviors of users, creates a system model of normal behavior with user's normal behavior and network data, and compares the difference the between detected data and the data in the normal behavior model so as to determine whether it is an attack. Misuse detection is by matching the intrusion to the signatures of known attacks. Most intrusion detection systems today adopt this approach.

With the rapid growth of the network information and the unlimited expansion of storage of information, how to analyze large amount of data processing effectively has become the bottleneck of intrusion detection system. Therefore, network intrusion detection technology must be able to adapt to high bandwidth and high load network environment and equipped a self-learning ability. Data mining technology has become the first choice of network intrusion.
K-means Clustering

Data mining is a process to extract potentially valuable knowledge (models or rules) from large amounts of data. It is a process using a variety of analysis tools to find the relationship between model and data in the mass data, which can be used to make predictions. Data mining tasks can be divided into two general categories: description and prediction[1]. Descriptive mining tasks characterize the general features of the database while predictive data mining tasks predict on the basis of the existing data.

**K-means Clustering Algorithm.** K-means algorithm is a widely used clustering algorithm. In K-means algorithm, k is the parameter, dividing n objects into k clusters for a high similarity within the cluster and low similarity between the clusters so as to classify the data. Algorithm first randomly select k objects as initial cluster centers. The rest objects, according to their distance from various clusters center, would be assigned to the nearest cluster. Then recalculate average number of each cluster and repeat the process until the criterion function is convergent[1].

The criterion function is Eq. 1:

\[
E = \sum_{i=1}^{k} \sum_{x \in C_i} |x - \overline{x_i}|^2
\]  

(1)

E is the sum of squared error of all data, \(x\) is a given data, \(\overline{x_i}\) is the average of the cluster. The distance use Euclidean distance, formula is Eq. 2:

\[
d(x, y) = \left[ \sum_{i=1}^{n} |x_i - y_i|^2 \right]^{1/2}
\]

(2)

The traditional k-means algorithm has the following disadvantages:

a. in K-means clustering algorithm, \(k\) should be given in advance. Given a set of samples, one may not know how many clusters are appropriate due to lack of experience or other reasons

b. in the k-means algorithm, you first need to determine an initial division based on the initial cluster centers. The choice of initial cluster center of cluster has great influence on the results. If the initial choice is not proper, one may not get clustering results effectively

c. the algorithm could only be used when average value of the cluster is given.

**Improvement of K-means Algorithm**

Because of the insufficiency of K-means algorithm, the choice of initial cluster centers and the calculation of the average value of cluster centers have been improved to some extent so that the clustering results have been improved.

**Improvement of the Selected Initial Cluster Centers.** Typically in a data space, high-density data object region is segmented by low-density object region. Usually points in the low-density region are noise points. In order to avoid getting the noise points, take \(k\) points of farthest distance in high density area as the initial cluster centers.

Define a density parameter to calculate the density region where the data object \(X_i\) is in: use \(X_i\) as the central, the density parameter is the radius of the data, expressed by \(\varepsilon\). \(\varepsilon\) is greater, the density of data is lower, otherwise, the density of data is higher. By calculating the density parameters of the data, the high-density data can be found, get a set \(D\) of high-density data. The distance between a point and a set is the closest distance of the point from the all points in the set.
In D, take the highest density region data object as the first Cluster center \( Z_1 \). Take a high-density point which has the farthest distance from \( Z_1 \) as the second Cluster center \( Z_2 \). Calculate the distance of the data \( X_i \) in D from \( Z_1 \) and \( Z_2 \), \( d(X_i, Z_1) \), \( d(X_i, Z_2) \), \( Z_3 \) is the \( X_i \) which is satisfy \( \max(\min(d(X_i, Z_1), d(X_i, Z_2))) \) (i=1,2,\ldots,n). \( Z_k \) is the \( X_i \) which is satisfy \( \max(\min(d(X_i, Z_1), d(X_i, Z_2), \ldots, d(X_i, Z_{k-1}))) \) (i=1,2,\ldots,n). So, \( k \) cluster centers can be found.

Specific process is as follows:

a. calculate the arbitrary distance between two data objects \( d(X_i, X_j) \).
b. calculate density parameter of each data object and delete the points in low-density regions to get data objects set \( D \) in high density regions.
c. take the data object in the highest density region as the first center \( Z_1 \), add it to the set \( Z \) and remove it from \( D \).
d. find the furthest point from \( Z \) in \( D \), add it to the set \( Z \) and remove it from \( D \).
e. Repeat d until the number of samples in \( Z \) reaches \( k \), i.e., find \( k \) initial cluster centers.

**Improvement of Algorithm with the Characteristics of Weighted.** In the data set which includes n data objects, each data object plays a different role in knowledge discovery. In order to distinguish the differences between them, each data object is assigned a weight. Here the weight setting method advanced by Domeniconi is adopted[3]. The basic principle of this method is to give greater weight for characteristics which has a good consistency within the cluster. Consistency in the distribution of cluster is measured of variance of the characteristics in cluster.

Suppose \( X \) represents the entire data set, \( X_i \) represents i class data set, \( X \) represents the data objects, \( E_{ir} \) represents i class variance of characteristics r, \( w_{ir} \) represents i class weight of characteristics r, \( c_k \) represents k class center vector.

\[
X_i = \{ x | j = \arg \min_k dist_w(c_k, x) \}
\]

\[
dist_w(c_k, x) = [\sum_{j=1}^{d} w_{ij}(c_{ij} - x_j)^2]^{1/2}
\]

\[
E_{ir} = [\sum_{x \in X_i} (c_{ir} - x_i)^2]^{1/2} / |X_i|
\]

\( c_{ir} \), represents the r characteristics of i class center, \( x_{ij} \), represents the j characteristics of data x, \( |X_i| \) represents the numbers of \( X_i \). \( w_{ir} \) is defined as Eq. 6:

\[
w_{ir} = \exp(-h \ast E_{ir}) / (\sum_{k=1}^{d} \exp(-2h \ast E_{ik}))^{1/2}
\]

\( h \) is a positive constant, defined as 12. The data objects need to be standardized first. In the experiments, it is found that better results can be achieved for \( x_j = x_j / \mu_{s_j} \), in which \( \mu_{s_j} \) is the average value of \( x_j \).

In summary, the improved algorithm process is as follows:

a. Choose \( k \) initial cluster centers with the above method, each object represents a cluster center.
b. set the initial weight \( \mathcal{W}_{ir} = 1 / d \), \( d \) represents the dimension of the data.
c. In accordance with the Eq. 3 and Eq. 4, divide each data objects into corresponding data object set. According to Eq. 5 and Eq. 6, calculate the new weight coefficients. 

\( \mathcal{W}_{ir} \)
d. According to the Eq. 3 and Eq. 4, recalculate the distance between data objects and the center of the clusters. Divide the data objects into the corresponding data object set.
e. Recalculate the cluster centers.
f. Repeat steps c, d, and e until the algorithm converges or reaches a certain number of iterations.

**Experimental Results and Analysis**

In this paper, KDD Cup 99 data packets is used in the intrusion detection experiment. KDD Cup 99 is the network data collected by simulating intrusion in the military network environment simulation, including nearly 5 million network connection records gathered by pre-treating TCP data frame. Each record includes categories of the normal behavior and aggressive behavior. Simulated attack data set can be divided into four categories: DoS (Denial of Service attacks), PROBE (scanning or other detection system), R2L (unauthorized access from remote computer), U2R (get super user’s rights unauthorized).

**Normalizing the Test Data**

Records of each Invasion in the packet are as follows:

```
0,tcp,http,SF,189,429,0,0,0,0,1,0,0,0,0,0,0,0,32,33,0,0,0,0,1,0,0,0,0,0,0,0,32,255,1,0,0,0,0,0,0,0,0,0,0,0,0.06,32,255,1,0,0,0.03,0.02,0,0,0,0,0,0,0,0.06,32,255,1,0,0,0.03,0.02,0,0,0,0,0,0,0,0.06,32,255,1,0,0,0.03,0.02,0,0,0,0,0,0,0,0.06,32,255,1,0,0,0.03,0.02,0,0,0,0,0,0,0,0.06,32,255,1,0,0,0.03,0.02,0,0,0,0,0,0,0,0.06,32,255,1,0,0,0.03,0.02,0,0,0,0,0,0,0,0.06,32,255,1,0,0,0.03,0.02,0,0,0,0,0,0,0,0.06
```

These records include 9 symbol property and 33 numeric property. In the experiment, we select 15 property to cluster. These values should be normalized. First, according to the Eq. 7 and Eq. 8, calculate the average value of each property \( m_f \) and the average value of absolute error \( s_f \).

\[
m_f = \frac{1}{n} \sum_{i=1}^{n} x_{if} \tag{7}
\]

\[
s_f = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_{if} - m_f)^2} \tag{8}
\]

\( m_f \) is the average of \( f \) attributes, \( s_f \) is the average error of \( f \) attributes, \( x_{if} \) is the \( f \) attributes of \( i \) record. Then use Eq. 9:

\[
z_f = \frac{x_{if} - m_f}{s_f} \tag{9}
\]

\( z_f \) is the \( f \) attributes after standardization. Then we complete normalization of test data.

**Experimental Results.** This algorithm is achieved under the environment of Windows XP and Visual C++6.0. Select 1000 normal data and 100 normal attack data from the invasion packets as the test data set M1. Select 10000 normal data and 1000 normal attack data from the invasion packets as the test data set M2. The number of clusters is 5. The results are divided into 5 categories. Results are shown as Table 1:
Table 1: The results of experiment

<table>
<thead>
<tr>
<th></th>
<th>The improved algorithm</th>
<th>k-means algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M1</td>
<td>M2</td>
</tr>
<tr>
<td><strong>AA</strong></td>
<td>93</td>
<td>990</td>
</tr>
<tr>
<td><strong>AN</strong></td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>detection rate of attack data</td>
<td>93%</td>
<td>99%</td>
</tr>
<tr>
<td><strong>NA</strong></td>
<td>8</td>
<td>178</td>
</tr>
<tr>
<td><strong>NN</strong></td>
<td>992</td>
<td>9822</td>
</tr>
<tr>
<td>detection rate of normal data</td>
<td>99.2%</td>
<td>98.22%</td>
</tr>
<tr>
<td>General detection rate</td>
<td>96.1%</td>
<td>98.61%</td>
</tr>
</tbody>
</table>

AA: attack data detected as the amount of attack data  
AN: attack data detected as the amount of normal data  
NA: normal data detected as the amount of attack data  
NN: normal data detected as the amount of normal data

**Conclusion**

With complexity of application software and operating system, network security is under increasing threat. Introducing data mining method to the network intrusion detection is beneficial in finding aggression and protecting the network security. On the basis of the traditional K-means, this paper adopts the improved K-means algorithm to the test network attack data, increasing the detection rate to some extent.

**References**