World Scientific Research Journal Volume 7 Issue 6, 2021
ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0015

Detection and Analysis of Network Intrusion Data Set Based on KNN Algorithm

Xiaojie Ma, Xiaorong Cheng

Department of Computer, North China Electric Power University, Baoding, Hebei, China

Abstract

With the continuous development of the network society and the frequent occurrence of network attacks, people's demand for network intrusion detection is increasing. The method of intrusion detection is basically to design a classifier that can distinguish the normal and abnormal data in the data stream, so as to realize the alarm of the attack behavior. This article will use the KDD99 data set in the academic circle to test the quality of intrusion detection algorithms to provide a unified performance evaluation benchmark for intrusion detection systems. This article will build a classifier based on the KNN algorithm, and use the 10% training set in the data set to train the classifier, and then use the corrected test set to test the classifier performance.

Keywords

Network intrusion detection; Data analysis; KNN algorithm.

1. INTRODUCTION

With the rapid development of computer network technology and the rapid increase of computer users, information security has become more and more important in computer networks [1]. In order to protect the security of computers and the Internet, traditional security measures of identity verification [2] and access control [3] have exposed many flaws and loopholes. The key to information and network security is intrusion detection.

Once the network intrusion detection system is designed and implemented, the key question is whether the intrusion detection system (IDS) meets its design goals. The method of evaluating intrusion detection system and the type of performance index required for evaluation are the problems to be solved in intrusion detection evaluation. The network penetration test system is a test software system with anti-mildew function under the above background and laboratory conditions. The system is mainly used to implement some common network attack methods. Intrusion detection system testing includes dividing known types of attacks into several categories, selecting some typical attack methods in each category, and simulating laboratory penetration tests [5]. The simulation results show that the system is feasible.

2. INTRUSION DETECTION DATA SET PREPROCESSING

In the KDD99 data set, each connection has 41 characteristics. Because some irrelevant or redundant features will reduce the quality, detection accuracy and speed of the classifier model, so the classifier extracting rules from many functions is time-consuming and inaccurate. In addition, the process of extracting features from raw tcpdump data is difficult and time-consuming, and may pose a fatal threat to online intrusion detection systems. Therefore, in order to improve the training speed and detection accuracy of the classifier, it is necessary to

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0015

remove redundant and irrelevant features. Different classifiers may have different optimal feature subsets.

This article uses 10% of KDD Cup 99 data to build a predictive model that can distinguish between intrusions or attacks and valuable connections. The database contains a set of standard data including various interventions simulated in the military network environment. KDD99 is divided into 4 types of attack types, followed by 39 sub-categories. The last item recorded in each row of the data set is used to mark the attack type. A total of 22 attack types appeared in the training set of this article, and the remaining 17 only appeared in the test set. The purpose of this design is to test the generalization ability of the classifier model. The detection ability of unknown attack types is good for evaluating the intrusion detection system. An important indicator of badness [6].

2.1. Numericalization of Data Sets

In data mining, the first task is data preprocessing, which aims to transform the data into machine-recognizable types and train them.

2.2. Data Set Standardization

In order to deal with the scattered data in the feature vector and big data swallowing small data, this article will standardize the data. In addition, data standardization can also shorten training time and suppress gradient explosions. This paper adopts Z-score standardization: the standardized distribution of the original data can be approximated to a Gaussian normal distribution with a mean of 0 and a variance of 1.

Assuming that the original data is X, the data mean is AVG, and the standard deviation is STD, the standardized formula is as follows:

$$X'_{ij} = \frac{X_{ij} - AVG_j}{STD_j} \tag{1}$$

$$AVG_j = \frac{1}{n}\left(X_{1j} + X_{2j} + \cdots + X_{nj}\right) \tag{2}$$

$$STD_j = \frac{1}{n}\left(\left|X_{1j} - AVG_j\right| + \left|X_{2j} - AVG_j\right| + \cdots + \left|X_{nj} - AVG_j\right|\right) \tag{3}$$

1) If $AVG_j = 0$, $X'_{ij} = 0$.

2) If $STD_j = 0$, $X'_{ij} = 0$.

2.3. Data Set Normalization

In many cases, different indicators have different dimensions and dimension units. This situation will distort the results of data analysis. Data standardization is needed to solve the comparability between data indicators in order to eliminate the influence of size between indicators. After the original data is standardized, the number of indicators is the same, which is suitable for comprehensive comparative evaluation.

Aiming at the linear transformation of the original data, this paper adopts the min-max standardization method to map the result value to [0,1]. Let max be the maximum value of the sample data, and min be the minimum value of the sample data. The conversion formula is:

$$x' = \frac{x - min}{max - min} \tag{4}$$

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0015

3. KNN ALGORITHM FOR DATA SET ANALYSIS

3.1. KNN Algorithm

The implementation principle of the KNN nearest neighbor classification algorithm: use all samples in the known category as a reference to determine the category of the unknown sample, and calculate the distance between the unknown sample and all known samples. Select the K known samples closest to the unknown sample, and classify the unknown sample and the K samples closest to it into one category according to the majority rule (majority decision).

The steps of the KNN algorithm:

(1) Data digitization

If the sample feature has a non-digital type, steps must be taken to quantify it as a number. For example, if the sample feature contains a color, you can calculate the distance by converting the color to a gray value.

(2) Data normalization

This example has multiple parameters, each of which has its own domain and value range. The impact on the distance calculation is different. For example, a larger value will affect a smaller value. Therefore, we need to scale the sample parameters. The easiest way is to standardize the values of all functions.

(3) A distance function is needed to calculate the distance between two samples

Commonly used distance functions are: Euclidean distance, cosine distance, Hamming distance, Manhattan distance, etc. Euclidean distance is usually selected as the distance metric, but it is only suitable for continuous variables. For discontinuous variables such as text classification, Hamming distance can be used as a metric. Generally, the use of certain special algorithms to calculate the metric can greatly improve the accuracy of the K-nearest neighbor classification, such as the use of the large-edge nearest neighbor method or the nearest neighbor component analysis method.

3.2. KNN Algorithm Implementation

```
def classify(input_vct, data_set):
    data_set_size = data_set.shape[0]
    diff_mat = np.tile(input_vct, (data_set_size, 1)) - data_set
    sq_diff_mat = diff_mat ** 2
    distance = sq_diff_mat.sum(axis=1) ** 0.5
    return distance.min(axis=0)
```

4. EXPERIMENTAL SIMULATION

The goal of the experiment is to actually test the network intrusion test system, and use the network intrusion test system to attack the target host under a well-built platform and environment to detect whether the system can achieve its due function. The experimental environment of this article: Windows 10, Python 3.8.

4.1. Precision Rate, Recall Rate, F Value

Suppose TP is a real example, FN is a false negative example, FP is a false positive example, and TN is a true negative example.

Precision: For a given test data set, the ratio of the predicted positive sample to the actual positive sample, the formula is as (5):

$$Precision = \frac{TP}{TP+FP} \tag{5}$$

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0015

Recall: For a given test data set, the ratio of the actual positive sample to the predicted positive sample, the formula is as (6):

$$Recall = \frac{TP}{TP+FN} \tag{6}$$

F1-score is used to balance the contradiction between the recall rate and the precision rate, that is, to select an appropriate balance point to consider the recall rate and the precision rate at the same time, the formula is as (7):

$$F - score = 2\frac{Precision \times Recall}{Precision+Recall} \tag{7}$$

Table 1. Calculates the accuracy rate, recall rate and F value of the 22 attack types in the intrusion detection system in the training set.

	Precision	Recall	F-score	support
0	1.00	1.00	1.00	38977
1	0.88	0.50	0.64	14
2	0.00	0.00	0.00	2
3	0.00	0.00	0.00	2
4	1.00	1.00	1.00	42797
5	1.00	1.00	1.00	112364
6	0.92	0.96	0.94	23
7	0.94	1.00	0.97	93
8	0.98	1.00	0.99	398
9	0.94	0.87	0.91	434
10	0.91	0.97	0.94	497
11	1.00	0.75	0.86	8
12	0.00	0.00	0.00	2
13	1.00	0.99	1.00	879
14	1.00	0.50	0.67	4
15	0.98	0.89	0.93	602
16	1.00	1.00	1.00	1
17	0.75	0.49	0.60	85
18	0.00	0.00	0.00	2
19	0.86	0.86	0.86	7
20	0.96	0.98	0.97	415
21	0.00	0.00	0.00	3

4.2. Normal and Attack Scatter Chart

Use Euclidean distance to calculate, and draw a scatter point distribution map. The horizontal axis is the serial number, and the vertical axis is the minimum Euclidean distance.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0015

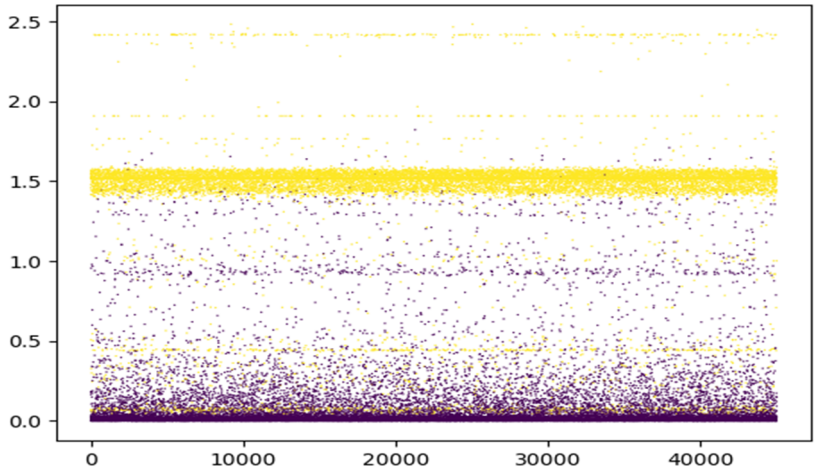

Figure 1. Scatter plot of normal and abnormal

4.3. ROC Curve Evaluation

TPR (True Positive Rate) represents the ratio of all samples that are actually positive that are correctly judged as positive, namely: TPR=TP/(TP+FN); FPR (False Positive Rate) means that all samples that are actually negative In the sample, the ratio of false positives, namely: FPR=FP/(FP+TN).

The ROC curve uses FPR as the X axis and TPR as the Y axis. The larger the FPR, the higher the degree of false reporting of the model, and the larger the TPR, the higher the degree of model prediction coverage.

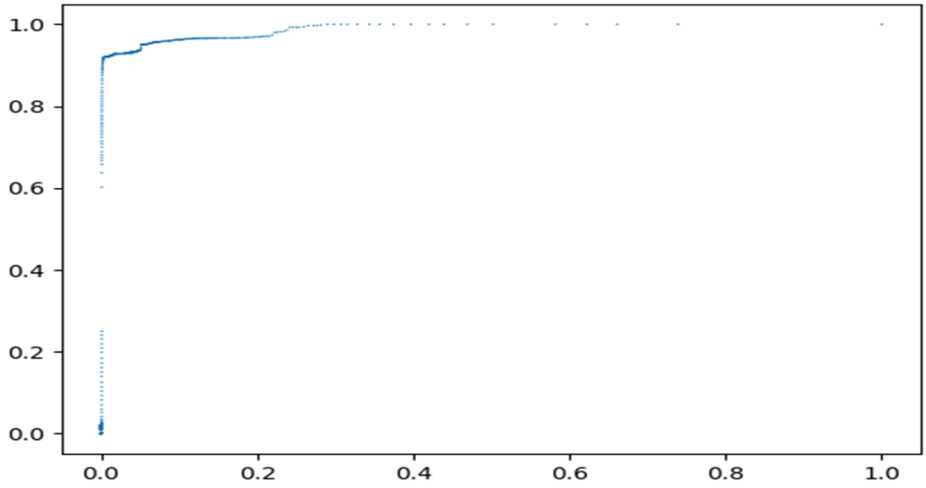

Figure 2. ROC curve

5. SUMMARY

In this paper, the character features in the data set are converted into numerical features, data standardization, data normalization, and combined with KNN algorithm to achieve intrusion detection data classification. Draw a scatter plot with serial number, minimum Euclidean distance, and class mark, and finally draw the ROC curve. K Neighbor (KNN) algorithm is a distance-based machine learning classification algorithm, which has the advantages of easy implementation, suitable for multiple classifications and fewer estimated parameters. But when the data is unbalanced, the performance of KNN is very poor.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0015

REFERENCES

[1] Liu Yuling, Feng Dengguo, Lian Yifeng, et al. Network security situation prediction method based on space-time dimension analysis[J].Computer Research and Development, 2014, 51 (8): 1681-1694.

[2] Zhao Wei. A design of WIFI system identity authentication based on network security [J]. Electronic Design Engineering, 2016, 24(14): 81-83.

[3] Fang Wenzhi.Network Security Access Control Technology[J]. Electronic Technology and Software Engineering, 2014(15): 212-213.

[4] Cui Congcong,Gong Shanshan.Basic Issues in the Establishment of the Global Cyber Security and Crime Convention[J]. Journal of Chongqing University of Posts and Telecommunications (Social Science Edition), 2015, 27(1): 23-28.

[5] Zhang Weihua.Design and Development of Network Intrusion Test System[J].Network Security Technology and Application, 2020(01): 11-14.

[6] Xu Feng. Research on intrusion detection model combining artificial neural network and genetic algorithm [D]. Jiangsu University of Science and Technology, 2015.