Cluster Analysis of Back Type of Denial of Service Attack

Jigang Zheng[1, a] and Jingmei Zhang[2, b]

[1]Department of Mathematic, Baoshan College, Baoshan, Yunnan, 678000, China.

[2]Library of Baoshan College, Baoshan, Yunnan, 678000, China.

[a]6913641@qq.com, [b]279619568@qq.com

Keywords: Intrusion detection; Weka; Characteristic property; Cluster analysis

Abstract. The characteristic attribute of network intrusion detection data set is an important index to measure the status of network intrusion. The analysis and research of the network intrusion detection system can deeply understand the current situation and the law of the network invasion. Clustering analysis using Weka software mining algorithm, the denial of service attacks "Back" type in accordance with the characteristics of the similarity of attributes is divided into 4 categories, and analyzes the characteristics of various types.

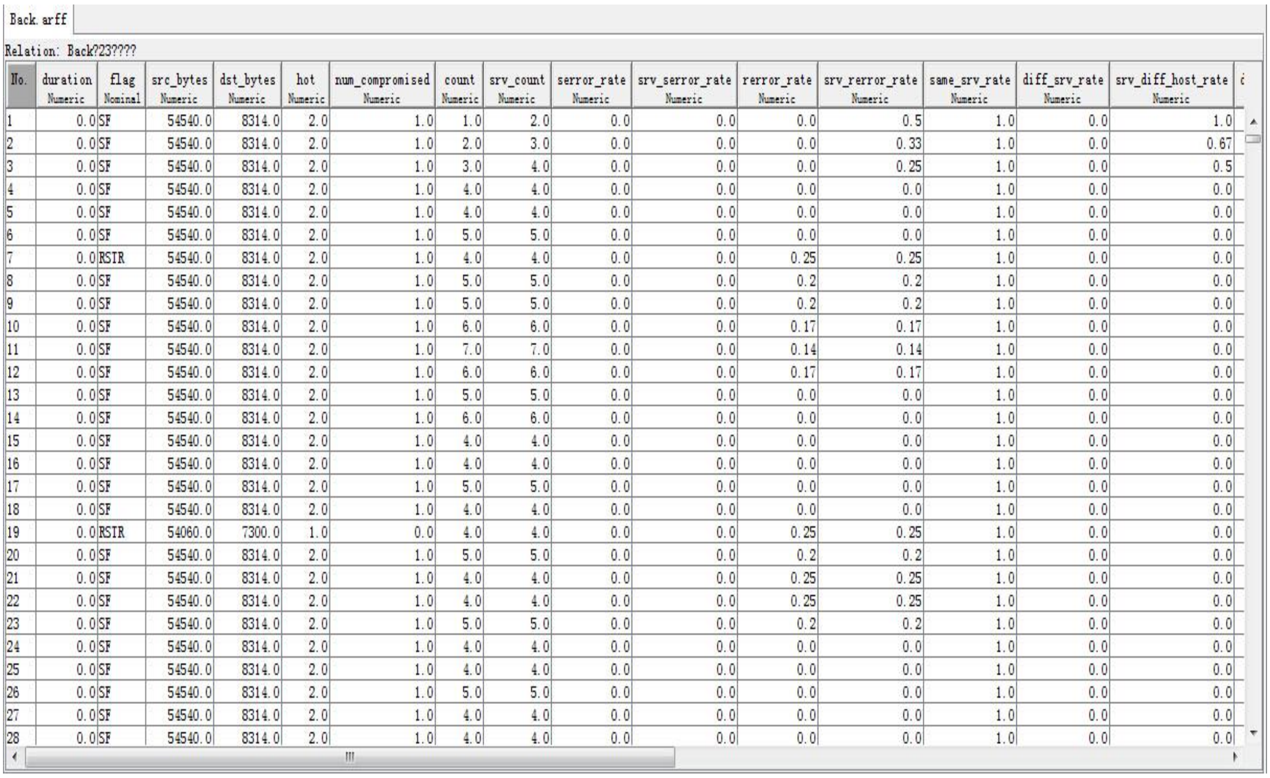
Introduction

Denial of service attacks (DoS) is currently widely used by hackers as a means of attack, it through exclusive network resources, so that other hosts cannot be accessed normally, resulting in network paralysis.[1]Famous operable to an intrusion detection data sets in data mining research network "KDDCup.data_10_percent 10 percent, the data set includes 494021 records, which has 391458 a denial of service attack records, accounting for the data set of more than.[2] 391458 denial of service attack records, the attack classification identified as "Back", "Land", "Neptune", "Pod", "Smurf", "Teardrop" and other six types of attacks. "Back" attack type has 2203, accounting for about 0.56% of total denial of service attacks. Weka's full name is the Waikato intelligence analysis environment (Waikato environment for knowledge analysis) is a Java based, for data mining and knowledge discovery of open source projects, the developers are from the University of Waikato in New Zealand, Ian H.Witten and Eibe Frank. After years of development, Weka is one of the most complete data mining tools, and is recognized as one of the most popular data mining open source projects. [3]

Mathematical modeling is to establish a mathematical model, to solve the actual problem of the process, is a pure mathematician into physicists, biologists, economists and even psychologists, and so on. Divided into three steps: building model, mathematical model, and model checking. Through the practical problems to find information, investigation, research its inherent characteristic and inherent laws, and put forward the necessary assumptions, built up to reflect the relationship between the number of the actual problem, for the model solution, and the model is applied to practical problems [1].At the same time for students to learn mathematics knowledge, computer knowledge and other aspects of the comprehensive knowledge, applied to practical problems, a reasonable explanation according to the calculation results consistent with the actual. China Undergraduate Mathematical Contest in modeling by the Higher Education Department of the Ministry of education and the Chinese society of industrial and applied mathematics, founded in 1992,the annual session. At present, it has become the largest based academic competitions, national organization of the Ministry of education of the college students four subject race one, is the world largest Mathematical Contest in modeling. In September 2015,about 80 thousand college students from 1326 universities in 33 provinces(city, autonomous region, including Hongkong and Macao) and Singapore signed up to participate in this competition.

Determine Data Mining Objectives

Each record of the KDDCup data set contains the first 41 fixed feature attributes and the last 1 attack types. The Back attack record has 19 attributes that are fixed and removed. That is protocol type is tcp, service is http, land is 0,wrong_fragment is 0,urgent is 0,num_failed_logins is 0,logged_in is 1,root_shell is 0,su_attempted is 0,num_root is 0,num_file_creations is 0,num_shells is 0,num_access_files is 0,num_outbound_cmds is 0,is_host_login is 0,is_guest_login is 0,dst_host_same_srv_rate is 1,dst_host_diff_srv_rate is 0,dst_host_srv_diff_host_rate is 0,23 feature attributes are left. ARFF format data in the Weka platform as shown in Fig. 1, through the Visualize all Preprocess visual interface, you can easily see the data classification summary visualization map, as shown in Fig. 2.



Back. arff

Relation: Back?23????

No.	duration	flag	src_bytes	dst_bytes	hot	num_compromised	count	srv_count	serror_rate	srv_serror_rate	rerror_rate	srv_rerror_rate	same_srv_rate	diff_srv_rate	srv_diff_host_rate
	Numeric	Nominal	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric
1	0.0	SF	54540.0	8314.0	2.0	1.0	1.0	2.0	0.0	0.0	0.0	0.5	1.0	0.0	1.0
2	0.0	SF	54540.0	8314.0	2.0	1.0	2.0	3.0	0.0	0.0	0.0	0.33	1.0	0.0	0.67
3	0.0	SF	54540.0	8314.0	2.0	1.0	3.0	4.0	0.0	0.0	0.0	0.25	1.0	0.0	0.5
4	0.0	SF	54540.0	8314.0	2.0	1.0	4.0	4.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
5	0.0	SF	54540.0	8314.0	2.0	1.0	4.0	4.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
6	0.0	SF	54540.0	8314.0	2.0	1.0	5.0	5.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
7	0.0	RSTR	54540.0	8314.0	2.0	1.0	4.0	4.0	0.0	0.0	0.25	0.25	1.0	0.0	0.0
8	0.0	SF	54540.0	8314.0	2.0	1.0	5.0	5.0	0.0	0.0	0.2	0.2	1.0	0.0	0.0
9	0.0	SF	54540.0	8314.0	2.0	1.0	5.0	5.0	0.0	0.0	0.2	0.2	1.0	0.0	0.0
10	0.0	SF	54540.0	8314.0	2.0	1.0	6.0	6.0	0.0	0.0	0.17	0.17	1.0	0.0	0.0
11	0.0	SF	54540.0	8314.0	2.0	1.0	7.0	7.0	0.0	0.0	0.14	0.14	1.0	0.0	0.0
12	0.0	SF	54540.0	8314.0	2.0	1.0	6.0	6.0	0.0	0.0	0.17	0.17	1.0	0.0	0.0
13	0.0	SF	54540.0	8314.0	2.0	1.0	5.0	5.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
14	0.0	SF	54540.0	8314.0	2.0	1.0	6.0	6.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
15	0.0	SF	54540.0	8314.0	2.0	1.0	4.0	4.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
16	0.0	SF	54540.0	8314.0	2.0	1.0	4.0	4.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
17	0.0	SF	54540.0	8314.0	2.0	1.0	5.0	5.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
18	0.0	SF	54540.0	8314.0	2.0	1.0	4.0	4.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
19	0.0	RSTR	54060.0	7300.0	1.0	0.0	4.0	4.0	0.0	0.0	0.25	0.25	1.0	0.0	0.0
20	0.0	SF	54540.0	8314.0	2.0	1.0	5.0	5.0	0.0	0.0	0.2	0.2	1.0	0.0	0.0
21	0.0	SF	54540.0	8314.0	2.0	1.0	4.0	4.0	0.0	0.0	0.25	0.25	1.0	0.0	0.0
22	0.0	SF	54540.0	8314.0	2.0	1.0	4.0	4.0	0.0	0.0	0.25	0.25	1.0	0.0	0.0
23	0.0	SF	54540.0	8314.0	2.0	1.0	5.0	5.0	0.0	0.0	0.2	0.2	1.0	0.0	0.0
24	0.0	SF	54540.0	8314.0	2.0	1.0	4.0	4.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
25	0.0	SF	54540.0	8314.0	2.0	1.0	4.0	4.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
26	0.0	SF	54540.0	8314.0	2.0	1.0	5.0	5.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
27	0.0	SF	54540.0	8314.0	2.0	1.0	4.0	4.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
28	0.0	SF	54540.0	8314.0	2.0	1.0	4.0	4.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0

Figure 1. Properties of ARFF format

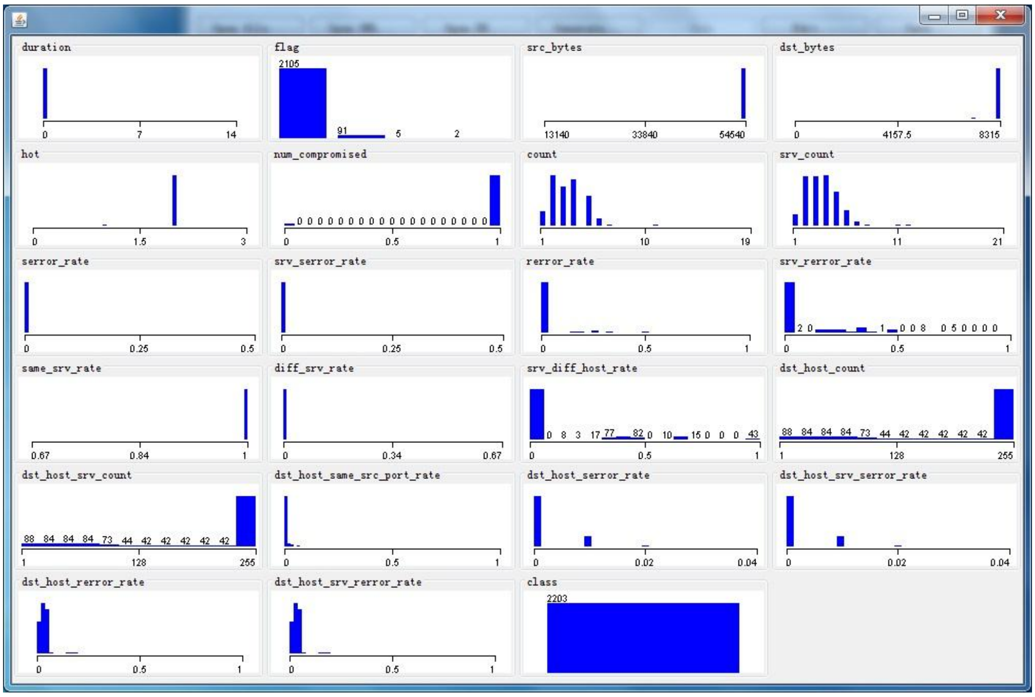


Figure 2. Visualization of attribute

Cluster Analysis

Clope,cobweb,DBSCAN,EM,FarthestFirst,FilteredClusterer,MakeDensityBasedCluster,optics,SIB, SimpleKMeans,XMeans a total of 11 kinds of clustering algorithm in Weka software, In this paper, the experimental analysis based on, algorithm SimpleKMeans on experimental data cluster.

SimpleKMeans algorithm referred to as the K-Means algorithm, K means algorithm, [4] described as follows:

Input: the number of clusters of K and a data set containing N data

Output: K cluster

Methods: K data were randomly selected as initial cluster centers

Random selection of K points as the initial centroid

Repeat

Assign each point to the nearest centroid, form K clusters, and calculate the centroid of each cluster from the new.

Until center of mass does not change

Select the appropriate random seed seed is key to effective clustering mining and enable the within cluster sum of squared errors "values the smaller the better. After several experiments, set parameters of 500 - s" weka.clusterer.SimpleKMeans - N 4 - I "weka.core.EuclideanDistance - R first-last" 5 "mining clustering results are shown in Table 1 below.

Table 1 Results of cluster analysis

Attribute	Full Data (2203)	Clustered0 (549)	Clustered1 (762)	Clustered2 (396)	Clustered3 (496)
duration	0.1289	0.0018	0.0052	0	0.5625
flag	SF	SF	SF	SF	SF
src_bytes	54156.3559	53941.0055	54540	54210.303	53762.2581
dst_bytes	8232.6496	8162.6412	8314	8235.3434	8183.0101
hot	1.9632	1.9107	2	1.9672	1.9617
num_compromised	0.9664	0.9235	1	0.9672	0.9617
count	3.3813	4.7195	2.4029	3.2146	3.5363
srv_count	3.6373	4.7413	2.4029	4.3485	3.744
serror_rate	0.003	0.0062	0.0032	0	0.0014
srv_serror_rate	0.0032	0.0062	0.0032	0	0.0024
rerror_rate	0.0405	0.0643	0.0012	0.061	0.0581
srv_rerror_rate	0.0958	0.0673	0.0012	0.3329	0.0835
same_srv_rate	0.9988	1	1	0.9983	0.9962
diff_srv_rate	0.0023	0	0	0.0034	0.0076
srv_diff_host_rate	0.1121	0.0063	0	0.5435	0.0572
dst_host_count	206.9637	249.2259	248.4331	246.1364	65.2016
dst_host_srv_count	206.9637	249.2259	248.4331	246.1364	65.2016
dst_host_same_src_port_rate	0.0103	0.0005	0.0006	0.0009	0.0435
dst_host_ser	0.0021	0.0003	0.0032	0.0022	0.0025

ror_rate					
dst_host_srv _serror_rate	0.0021	0.0003	0.0032	0.0022	0.0025
dst_host_rer ror_rate	0.0504	0.0444	0.0353	0.0392	0.0893
dst_host_srv _rerror_rate	0.0504	0.0444	0.0353	0.0392	0.0893
class	back.	back.	back.	back.	back.

The clustering results into 4 categories,Clustered0 said the connection time is 0.0018 seconds and the length of duration is SF, flag connected the source host to the destination host src_bytes data flow is 53941.0055B,the destination host to the source host dst_bytes data flow is 8162.6412B,the number of hot indicator is 1.9107 hot, threatened by the number num_compromised for 0.9235,the same target host connection number count is 4.7195,the same service connection number is 4.7413 srv_count, with a host of SYN error serror_rate accounted for 0.62%,the same service SYN wrong srv_serror_rate accounted for 0.62%,with a host of REJ error rerror_rate accounted for 6.43%,with a REJ service error srv_rerror_rate the proportion of 6.73%,with a host of the same service same_srv_rate accounted for 100%,the same host not Occupied error occupied connection rej occupied error occupied the DST SRC host port for connection 249.2259 occupied the same service diff SRV rate ratio was 0%,with a different host SRV diff host rate accounted for the proportion of 0.63%,and the connection with the same target host connections DST host count is the same service quantity of DST host SRV count 249.2259,same source DST same port rate accounted for the proportion of 0.05%,syn wrong connection host Dst_host_serror_rate rate ratio of 0.03%,the same service syn DST host SRV Dst_host_srv_serror_rate rate ratio of 0.03%,the error of the DST host Dst_host_rerror_rate rate ratio of 4.44%,same service rej DST host SRV Dst_host_srv_rerror_rate rate ratio of 4.44%.The clustering results indicated by Clustered1,Clustered2,Clustered3,Clustered4 and Clustered5 are clearly presented in Table 1.

Conclusion

With the help of the software Weka3.6.13 version of the famous open source data mining, the KDDCUP99 data set "KDDCUP.data_10_percent 10 percent concentrated refuse service attack" back "type cluster analysis was carried out, 2203 intrusion data records according to the similarity of attributes is divided into four categories, the network intrusion detection data set inherent law have a certain understanding, analyzed the intrusion data records.

Reference

[1] Xi Lei, Wang Feng, Wei Xiuran, Yu Hua, Zhang Hao. Design of Host-based Defense System against Smurf Attack[J].Joumal of North China Institute of Water Conservancy and Hydroelectric Power,2006(2):66-68

[2] University of California. KDD Cup 1999 Data [EB/OL]

http://kdd.ics.uci.edu/databases/KDDCUP99/KDDCUP99.html,1999-10-28.

[3] Wang Xuehui, Jia Lili. Weka Makes Data Mining no Longer be Mystical [J].Computer Knowledge and Technology, 2007(5):699.

[4] Pang-Ning Tan, Michael Steinbach, Vipin Kumar. Data Mining Introduction. Bei Jing: The people post and Telecommunications Press, 2006:310.