The Research about Data Mining of Network Intrusion Based on Apriori Algorithm

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

[1]Department of Mathmatic, Baoshan College, Baoshan, Yunnan, 678000, China

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

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

Keywords: Apriori algorithm; Data mining; Network intrusion; Association rules

Abstract. Found that the association between each characteristic attributes and behavior based on network intrusion data to achieve intrusion valuable data extraction and analysis, the use of denial of service attacks recorded KDDCup99 dataset simulation experiments. Weak software using association rule mining algorithms for different types of denial of service attack attribute characteristics for analysis, has been contact between the different characteristics of different attributes classification identifies the types of attacks, to improve the efficiency and accuracy of intrusion detection has an excellent role, has a certain value.

Introduction

With the rapid growth of network bandwidth, hacker attacks are becoming increasingly diverse, the use of computer network crime shows a clear upward trend. How to establish a secure and network system to ensure the security of important information has become an urgent problem. The existing intrusion detection system has slow response and poor real-time performance when the network is invaded. How to deal with the massive data in real-time and discover the attack in time will become the key problem in the next research of intrusion detection system.

The network intrusion detection is firewall, it can prevent the use of protocol, source routing, address counterfeiting and other means of attack, and provide secure data channel, but it is for the application layer of the back door, internal users unauthorized operation as a result of the attack or steal, destroy the information but incapable of action. In addition, due to the location in the firewall in the network open, design defects will inevitably be exposed to the attacker, so only by virtue of the firewall is difficult to resist various attacks emerge in an endless stream.

Data mining is one of important research subjects, it has not yet been found to dig out the hidden and useful, information and knowledge from a large number of original data, help the decision-maker to find useful knowledge from the data between the potential, it has been received wide attention in the field of research and application. In the intrusion detection system in the rational use of data mining technology, the user behavior characteristics can be extracted through the analysis of historical data, summarize the law of intrusion, so as to establish a complete rule base for intrusion detection. Apriori algorithm is a classical algorithm for discovering association rules in data mining. The concepts of association rule mining and Apriori algorithm are proposed by Agrawal et al[1].Used to find interesting association rules or related relationships among data items in a given dataset, and a mesh relation graph between association rules and feature attributes and intrusion types. Through the analysis of data mining association rules mining the potential of information and data from the historical statistics show the character of behavior, behavior, will these to the intrusion detection database, the statistical characteristics of the current user behavior and historical data are compared, and the security policy conflict behavior is determined as the intrusion behavior[2].

Apriori Algorithm

Algorithm Overview. The Apriori algorithm first finds all frequent itemsets, and then generate strong association rules from the frequent item sets. The first step is to find all the frequent item support is not less than the user specified minimum support threshold from the transaction database in D; the second step, the use of frequent item sets of association rules produce the desired, the basic principle of generating association the rules of the confidence must not be less than the minimum confidence threshold specified by the user. Because the second step is easy and intuitive, the first step mining all the frequent item sets is the core of the algorithm, occupying the whole calculation of the majority, so sometimes only consider the efficiency of mining frequent item sets.

Definition1[3] Set $I=\{I_1,I_2,\cdots,I_m\}$ is a collection of data items, D transaction is a collection of all, A transaction T has a unique identifier TID.If items, transaction support items claimed T set A,also known as T transaction that contains the item set A.

Definition2 Association rules are shaped like $A\Rightarrow B$ type of implication, among them $A\subset I$, $B\subset I$,and $A\cap B=\Phi$.

Definition3 $A\Rightarrow B$ supports the association rule is defined as: $\sup port(A\Rightarrow B)=\frac{\sup port(A\cup B)}{N}\times 100\%$,credibility is defined as: $confidence(A\Rightarrow B)=\frac{\sup port(A\cup B)}{\sup port(A)}\times 100\%$

Algorithm Description. In:Database D and $\min\_\sup$,

Out: Database D item sets L,

Algorithm:

L_1= Looking frequent two sets(D);

For k=2;$L_{k-1}\neq\Phi$; k++

{ C_k=apriori_gen(L_{k-1});

For each transaction $t\in D$

{ $C_t=subset(C_k,t)$;

For each candidate $c\in C_t$

$c.count++$;}

L_k=($c\in C_k\mid c.count\geq\min\_\sup$)}

Return L={All L_k}.

The apriori_gen is a key step in Apriori algorithm, according to the L_{k-1} for L_k, need to do two things: pruning and connection. The connection step: to produce C_k, by connecting the pruning step: if a candidate k set ($k-1$) a subset of frequent item sets in ($k-1$),then the candidate set is not frequent, so as to remove from C_k [4].

Apriori_gen is described as follows:

apriori_gen (L_{k-1}:frequent(k-1)-itemsets)

For each item sets $l_1\in L_{k-1}$

For each item sets $l_2\in L_{k-1}$

If $(l_1[1]=l_2[1])\wedge\ldots\wedge(l_1[k-2]=l_2[k-2])\wedge(l_1[k-1]<l_2[k-1])$

Then

{ $c = l_1 \, l_2$;
If has_infrequent_subset (c, L_{k-1})
Then delete;
Else add c to C_k;
}
Return C_k.

Simulation Experiment

Experimental environment: Intel Pentium (R) 3.0GHz main frequency,2G memory,500G hard disk; 64 bit Microsoft Windows 7 operating system; algorithm using Java7.0,Weka3.6.2 implementation.

Data Set Selection. The reference data from the classic KDDCup99 intrusion detection data sets, a total of 4898345 data records in the data set, the normal data of 972780,covering 4 types of attacks: Probing41 102,R2L90,DoS3 883370,U2R1 3[5].We choose KDDCup99 10% training data set, the data set of a total of 494021 records, including Service denial of service attacks (Denial Of, DoS) records, accounting for about 79.24% of the data set[6].Each record contains the attributes of fixed the first 41 and the last 1 attack types identification feature, select the first 13 attributes of the duration,protocol_type,service,flag,src_bytes,dst_bytes,land,wrong_fragment,urgent,hot,num_failed _logins,logged_in,num_compromised and mark the last 1 types of attack attributes class attack classification of back, land, identified as Neptune, pod, Smurf, teardrop six types of attacks, removes the influence of the mining results is not the other 28 attributes, visual attributes as shown in Fig. 1,the numerical characteristics of the discretization of continuous attributes for the classification attribute.

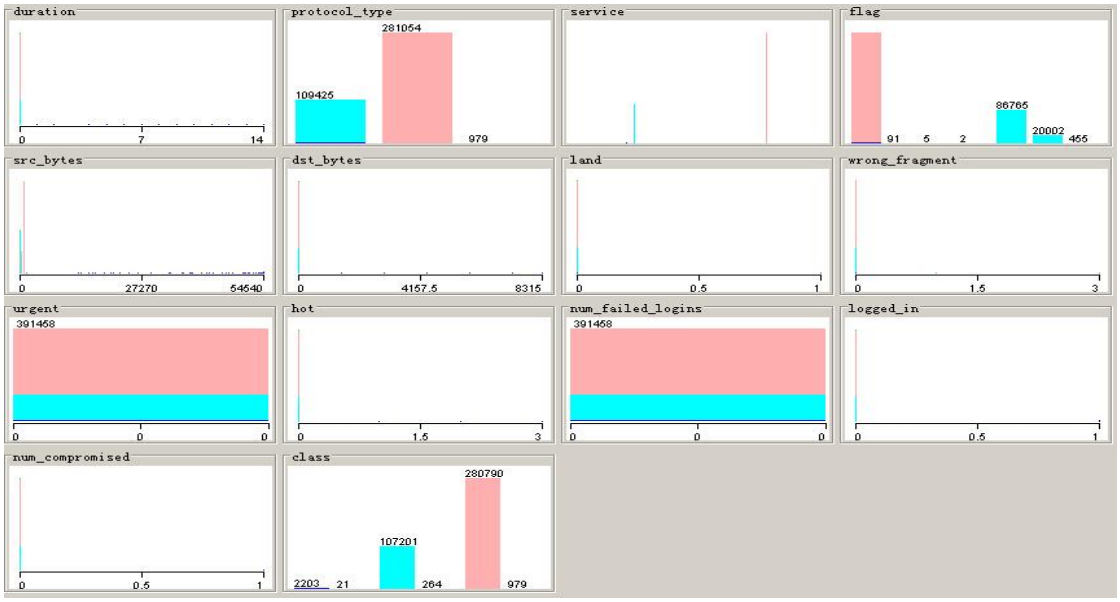

Figure 1. The 14 attributes of visualization

Association Rule Mining. Select the appropriate minimum support and minimum confidence, partial mining results are as follows:

1.num_compromised=1,dst_bytes=[1506-9004],src_bytes>=1032,service=http,hot=2,flag=SF==>class=back

2.service=nnsp, flag=S0==>class= neptune

3.dst_host_serror_rate=(-inf-0.13],dst_host_rerror_rate=(-inf-0.206667],dst_host_diff_srv_rate=(-inf-0.086667],dst_host_count=(229.8-inf)==>class=smurf

4.service=private, flag=SF==>class=teardrop.

According to the above mining results, forced to compromise appearance number of num_compromised is 1, the destination host to the source host data flow dst_bytes between 1506 and 9004 bytes, the src_bytes data flow source host to the destination host more than 1032 bytes, the network service type service destination host HTTP, access to sensitive system files and directory number hot 2, flag connection state normal or wrong SF, attack classification mark class as back. Destination host network service type service is NNSP, connecting normal or error status flag to S0, attack type identifier class is neptune. SYN the percentage dst_host_serror_rate connection error is less than 13% REJ, the percentage dst_host_rerror_rate connection error is less than 20.67%,and the connection with the percentage of dst_host_diff_srv_rate is less than 8.67% with the same target host of different services, and the connection number dst_host_count with the same target host is greater than 230,the type of attack classification identification class Smurf, the network service type service destination host private, flag connection state normal or wrong SF, attack classification mark class as teardrop. Reduce the minimum support and minimum confidence, mining more association rules.

Can also use the improved Apriori algorithm of data mining for data sets with good properties, the improved Apriori algorithm to optimize the number of candidate sets only, reduce the amount of calculation of frequent itemsets, and improve the efficiency of database scanning, scan the transaction database only contains frequent item sets[7].

Conclusion

Apriori algorithm is used for data mining of denial of service attack data set of network intrusion data. The data mining software Weka3.6.2 version has good mining effect. According to the relationship between the characteristic attributes and behaviors, the network intrusion detection system is developed.

Acknowledgements

This paper is the 2016 stage of the Baoshan College Institute for scientific research Project" Research on improvement and application of association rule mining algorithm "(No.2016BY007).

References

[1] Agrawal R, Srikant R.Fast Algorithm for Mining Association Rules. In Proceeding 1994 International conference Very Large Data Base(VLDB’94).Santiago,Chile,Sept,1994:487-499.

[2] Wang Zhongcai, Li Yongbi. Intrusion Detection System Based on Data Mining Research[J].Bulletin of Science and Technology,2012,(8):150-152.

[3] Pang-Ning Tan, Michael Steinbach, Vipin Kumar. Data Mining Introduction. Bei Jing: The people post and Telecommunications Press, 2006:201-204.

[4] Jiawei Han, Micheline Kamber. Data Mining Concepts and Techniques [M].Beijing: Machinery Industry Press, 2007:151-154.

[5] Chen Zhen. Research on the Intrusion Detection Systems Based on the Improved Apriori Algorithm[J].Journal of Hainan Normal University (Natural Science),2012,(1):41-45.

[6] Zhang Xinyou, ZENG Huashen, JIA Lei. Research of intrusion detection system dataset KDD CUP99 [J].Computer Engineering and Design, 2010, 31(22):4809-4816.

[7] Ding Li. The Research about Data Mining of User's Behaviors Based on Apriori Algorithm[J].Bulletin of Science and Technology,2013,(12):214-217.