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

Jigang ZHENG[1, a],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

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.Weka 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.

Keywords: Apriori algorithm;data mining;network intrusion;association rules;denial of service

Introduction

With the rapid growth of network bandwidth,hacker attacks increasingly diverse,the use of computer network crime showing a clear upward trend.How to build safe and network system to ensure the security of critical information has become a serious problem.Existing network intrusion detection system in the invasion,the reaction is slow,real poor.How to deal with real-time network,vast amounts of data and the timely discovery of the attack will be the focus of the next issue of intrusion detection systems research.

Network intrusion detection methods are mostly used in the past the firewall policy,it can prevent the use of protocol vulnerabilities,source routing,various methods to address phishing and other attacks,and provide secure data channel,but it is for the back door of the application layer,internal users unauthorized operation as a result of an attack or theft,destruction of information are powerless.In addition,due to the location in the network firewall daylight,their design flaws will inevitably be exposed to a large number of attackers,it is only by virtue of a firewall is difficult to resist the endless variety of attacks.

Data mining is one of the important research topic,it dig out from a lot of raw data implied,useful information and knowledge not yet been found to help decision-makers to find among the data potentially useful knowledge from a proposal to now has been wide attention and research applications.Rational use of data mining techniques in intrusion detection system,by analyzing the historical data can be extracted behavioral characteristics of users,summed up the law intrusion,in order to establish a more complete rule base for intrusion detection.Apriori algorithm is a data mining association rules found in classical algorithm proposed by Agrawal and other concepts related to association rule mining and Apriori algorithm[1],For looking for interesting association rules or correlation between a given data set of data items,as well as the rules and feature attributes associated with the type of network connection diagram invasion.By association rules data mining to analyze the potential of information and data that can be tapped from the historical behavior,showing statistical characteristics of behavior among these will be added to a database intrusion detection,the statistical properties of the current user behavior and historical data are compared with each other and security policy contradictory behavior is determined that the intrusion[2].

1 Apriori algorithm

1.1 Outline

Apriori algorithm first find all the frequent item sets,and then generates a strong association rules from frequent itemsets: First step,frequent item sets to find out from the transaction database D in support of not less than all of the user-specified $\min\_\sup$ threshold;The second step,the basic principles of the use of frequent item sets to produce the desired association rules,association rules is to generate confidence must not be less than the user-specified $\min\_conf$ threshold total.The second step is more due to easy and intuitive,so the first step to dig out all frequent item sets is the core of the algorithm,occupy most of the entire amount of computation,it is sometimes only consider the efficiency of mining frequent item sets.

1.2 Description

In:Database D and $\min\_\sup$,

Out:Database D itemsets 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 $L_k$}.
```

apriori_gen is a critical step of Apriori algorithm,According L_{k-1} to find L_k,need to do two actions:connection and pruning.By connecting generation C_k,If a candidate itemsets k subset ($k-1$) of ($k-1$) is not focused on frequent item,the candidate set and can not be frequent,thus deleted by C_k[3].

apriori_gen description follows:

```
apriori_gen($L_{k-1}$:frequent (k-1) itemsets)
    For each itemset $l_1 \in L_{k-1}$
        For each itemset $l_2 \in L_{k-1}$
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 $C_k$.
```

Among them:

l_1[1]=l_2[1]∧l_1[2]=l_2[2] ∧l_1[k-2]≠∧l_2[k-2] ∧l_1[k-1]<l_2[k-1]

2 Simulation

Experimental environment:Intel Pentium(R) G2030 3.0GHz,2G RAM,500G hard disk;Microsoft Windows 7;Java7.0,Weka3.6.2.

2.1 Dataset

Experimental reference data from KDDCup99,there are 4898345 data records,in which 972780 normal data,covering four types of attack:Probing 41102 records,R2L90 records,DoS 3883370 records,U2R 1003 records[4].Here we choose KDDCup99 10% training data set,the data set record total of 494021,of which 391458 denial of service attacks (Denial Of Service,DoS) record,accounting for 79.24 percent of the data set[5].Each record contains the top 41 fixed feature attributes and finally a characteristic attribute identifies the type of attack,Select property before 13 features:duration、protocol_type、service、flag、src_bytes、dst_bytes、land、wrong_fragment、urgent、hot、num_failed_logins、logged_in、num_compromised and finally an attack attribute identifies the type of feature class,classification identifies the types of attacks:back、land、neptune、pod、smurf、teardrop,removes little impact on the mining results remaining 28 feature attributes,characteristic properties visualized as shown in Figure 1,characterized by continuous numeric attributes into discrete classification characteristic properties.

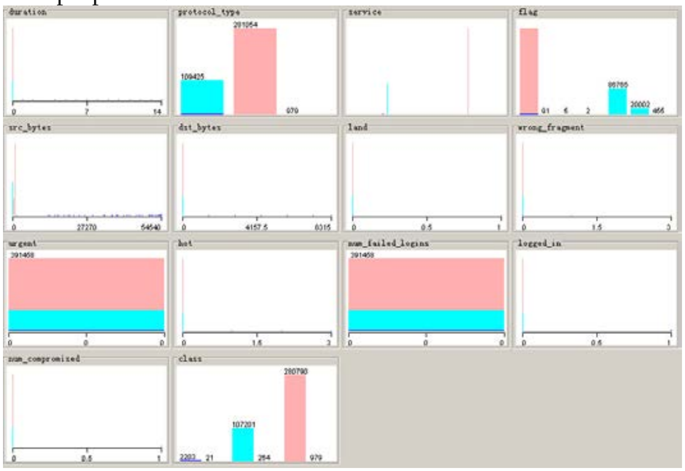

Fig.1 The 14 attributes of visualization

2.2 Association rule mining

Select the appropriate minimum support and minimum confidence,some mining results 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,the number of occurrences num_compromised forced to compromise for a destination host to the source host dst_bytes data traffic between 1506-9004 bytes,the source host to the destination host src_bytes greater than 1032 bytes of data traffic,the destination host the network service type service for http,the number of access to the system of sensitive files and directories hot for two times,the connection is normal or error status flag for SF,attack type classification identifies the class is back.The purpose of the host's network service type service for nnsp,the connection is normal or error status flag for S0,attack type classification identifies the class of neptune.SYN connection error occurs percentage dst_host_serror_rate less than 13%,there REJ wrong connection dst_host_rerror_rate percentage share of less than 20.67%,the connection with the current percentage dst_host_diff_srv_rate different services with the same target host connections share is less than 8.67%,and the current connection with the same number of connections dst_host_count target host more than 230 types of attacks classification identifies the class of smurf,the destination host's network service type service is private,the connection is normal or error status flag for SF,attack type classification identifies the class as a teardrop.Reduce the minimum support and minimum confidence,get more mining association rules.

3 Summary

Apriori algorithm uses the data network intrusion denial of service attack datasets for data mining,data mining software Weka3.6.2 version,with good mining results.According to the association this article excavated characteristic attributes and behaviors,and then develop a network intrusion detection systems,has good application prospects.

Reference

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

[4] 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.

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