Research and Improvement on ID3 Algorithm in Intrusion Detection System

Guangqun Zhai
School of Information Engineering
Zhengzhou University
Zhengzhou, China

Chunyan Liu
School of Information Engineering
Zhengzhou University
Zhengzhou, China

***Abstract*—ID3 algorithm was a classic classification of data mining. It always selected the attribute with many values. The attribute with many values wasn't the correct one, and it always created wrong classification. In the application of intrusion detection system, it would created fault alarm and omission alarm. To this fault, an improved decision tree algorithm was proposed. Though improvement of information gain formula, the correct attribute would be got. The decision tree was created after the data collected classified correctly. The tree would be not high and has a few of branches. The rule set would be got based on the decision tree. Experimental results showed the effectiveness of the algorithm, false alarm rate and omission rate decreased, increasing the detection rate and reducing the space consumption.**

Keywords- ID3 algorithm; information entropy; information gain; rule; intrusion detection

I. INTRODUCTION

With the development of network attacks technology, risk prevention of network security has kept rising. The original defensive measures can not catch the intrusion events. In order to guarantee the security of network information better, Intrusion Detection System (IDS) was proposed. After Firewall, the IDS has been the second security line. It could recognize intrusion events and watch successful safe break, but it has high false alarm and omission alarm.

Now, many technologies of data mining are used in IDS, for example, decision tree, GA algorithm and immunology. In IDS, data analysis was very important. A good data analysis has high detection speed, low false alarm. The quality of data analysis is related directly to the safe of information. There were many data analysis methods. To some extant, they could find the intrusion events, but they could not find the changed intrusion. IDS has high false alarm and omission alarm. The data analysis methods need to be improved, new algorithm need to be proposed. Processing the data collected is to determine the intrusion event. At first, the dada collected was classified. The decision tree was one of classification methods which had good effects. ID3 algorithm was important algorithm to decision tree.

II. ID3 ALGORITHM

ID3 algorithm [1] is a top-down and greedy inductive learning methods. Its core idea is information entropy theory. Selecting the properties of the largest information gain as the categorical attribute, expanding the decision tree branches recursively, at last, the decision tree [2, 3, 4] is created.

The set S has s pieces of records data samples. Assumed that the data packet has m different attribute values, defining m different types C_i (i=1, 2, …, m). Assumed that the class C_i has s samples. For a given sample, the expectations of the information required for classification is given by:

$$I\left(s_1, s_2, \ldots, s_m\right) = -\sum_{i=1}^{m} Pi \log_2(Pi) \tag{1}$$

In this formula, the value P_i is the probability of an arbitrary sample belongs to C_i, and the value was estimated s_i / S.

Assumed that the attribute A has v different values $\{a_1, a_2, \cdots, a_v\}$. Using attribute A divides set S into v subsets. In them, the s_i includes the samples, which the value of attribute A is a_i in set S. If A is selected as test property (Namely, the best splitting attribute), these subsets includes all branches, which were brought about by nodes of set S. Assumed that s_{ij} is the number of samples, subset s_i is belong to C_i. The entropy of subnets classified by A is:

$$E(A) = \sum \frac{s_{1j} + s_{2j} + \ldots + s_{mj}}{S} I\left(s_{1j}, s_{2j}, \ldots, s_{mj}\right) \tag{2}$$

In this formula, item $(s_{1j}+s_{2j}+\ldots+s_{mj})/S$ is seen as the weight of subset j, and the value is number of samples divided number of set S. Expectation information of given subset is:

$$I\left(s_{1j}, s_{2j}, \ldots, s_{mj}\right) = -\sum_{i=1}^{m} p_{ij} \log_2\left(p_{ij}\right) \tag{3}$$

In this formula, the value $p_{ij} = s_{ij}/|S|$ is the probability of the sample belong to class C_i. A branch on attribute A will get the information in the following:

$$Gain(A) = I\left(s_{1j}, s_{2j}, \ldots, s_{mj}\right) - E(A) \tag{4}$$

The information gain of each attribute is calculated. The attribute is the tests attribute of set S, which has the biggest information gain. Creating a node, using its attribute as a mark,

FUND: Scientific and technological project of Henan Province (0423020300)

creating branches dependence on value of each attribute, and then dividing data sample.

III. SHORTCOMINGS OF ID3 ALGORITHM AND ITS IMPROVEMENT

A. Application of ID3 Algorithm in IDS

Intrusion detection system is used to identify illegal attacks of computer systems and network. The main purpose of IDS is to protect their information resources. It is different from other security products. It needs more intelligence. Dividing data collected into categories, getting the useful results. The method attribute analysis is used commonly in data analysis.

ID3 algorithm chooses the attributes with maximum gain(A) as the root node[1]. The key point of using ID3 algorithm is to fix the attribute of classification. If the attribute selected is not suitable, the decision tree will have more nodes and longer path. It will result in that the rules set may have redundancy. When classifying data, a good attribute should be selected.

Decision tree is one of example learning methods, it can be used individually as a "pre-processing module". Classifying intrusion pattern, choosing attack attribute as a sample, then a decision tree will be created though attribute classification. A path from the root node to one leaf node is a piece of rule. Combined with pattern matching algorithm and rule set, the intrusion event would be detected.

B. Shortcomings of ID3 Algorithm

Because of the diversity of attacks, when collecting patterns characteristic and processing data, if the data sample is divided by original ID3 algorithm, decision tree will have more Redundancies. In this case, it will result in wrong message when the input event matches rule base.

The shortcomings of ID3 algorithm can be described in the following: Assumed that there is attribute named A. It has n different values (v_1, v_2, v_3, ⋯, v_n). If there is a new attribute A′, it has n+1 different values (v_1, v_2, v_3, ⋯, v_{n-1}, v_{n+1}, v_{n+2}). A′ is different form A. The value v_n is changed into values v_{n+1} and v_{n+2}, the other n-1 values are completely same. Assumed that the data sample had been classified correctly by attribute A. Usually, ID3 algorithm chooses the attributes with maximum Gain(A) as the root node, in this case, attribute A′ must be selected. Practically, attribute A′ is not better than attribute A. It will play the opposite role. To these shortcomings, it needs to improve ID3 algorithm [5, 6, 7].

C. Improvement of ID3 Algorithm

The standard of selecting classification attribute is the attribute has maximum gain(A). Information entropy [1] is a chaotic level statistics. The bigger the information entropy is, the more excursive the system is. The purpose of classification is drawing system messages, and making the system order, rule, organized. The information gain is the reduction of segregation entropy. The standard of classification is selecting the attribute with maximum gain(A).

At first, the formula of information gain is modified (the 4th formula). The new information gain formula is:

$$Gain'(A) = Gain(A) * (n - m) \tag{5}$$

In this formula, the number of sample is n, and m is the number of values of attribute A. When selecting attribute, the attribute with maximum information gain replaces of original attribute. Though this improvement, correct attribute would be chosen. A new decision tree will be built based on the new formula.

The data sample is captured from experiment. Two decision trees will be created. One is created by original algorithm, and the other one is by improved algorithm. Comparing the two trees, the effectiveness of improvement is proved.

Assumed that T is the training sample set, and T_attribute_list is the candidate attribute set. The improved ID3 algorithm can be described in the following lines.

Input: training sample set T, candidate attribute set T_attribute_list

Output: a decision tree

1) Create root node R.

2) If T is belongs to the same class C, return R as a leaf node, and mark it class C.

3) If T_attribute_list is empty, return: R is a leaf node, and marks the class with most occurrences in R.

4) FOR each T_attributelist, calculate the information gain by improved ID3 algorithm, select the attribute with bigger values as the classified node. If the values of information gain are equal, select the one with small order number.

5) FOR different values of T_attribute_list {

A new leaf node is drawn out from R node;

IF the new leaf node is belongs to T', and T' is empty, not divide the node, mark it with most occurrences in T.

ELSE

Run Create_decision_tree (T′, T′_attribute_list) on the node, go on to divide it.}

IV. EXPERIMENT AND DATA ANALYSIS

A. Simulated Experiment

Snort [8] is a cross-platform and lightweight detection troop. It can be used to watch small TCP/IP network and detect all kinds of suspicious network traffic. Fig. 1 shows us the workflow of Snort. The system calls Winpcap library function, capturing packet from network, and then calling packet analytic function. After the packet is pretreatment, the system will start detection engine, matching decoding packet data and two-dimensional rule list. If the matching succeeds, it means intrusion happened. According to the provision ways, the system responds alarms or response, after this, the process of this packet is over. If matching is failure, there is no intrusion, return directly. The system goes on to capture next packet.

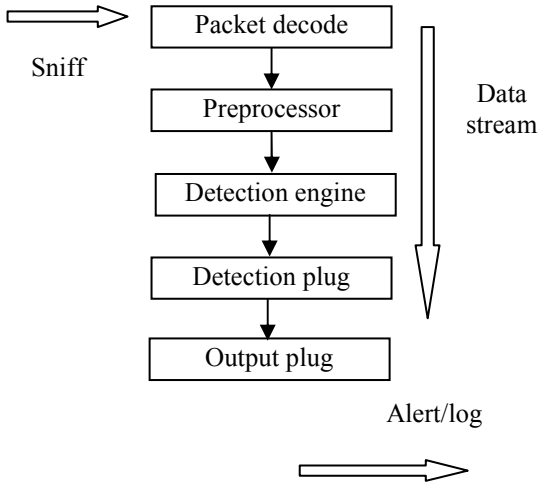

Figure 1. Workflow of Snort

Rules [9] are the core of Snort. After the rules are formatted rule list, searching and matching are easy. The rule of Snort is divided into two parts: head and option. The head includes action, protocol, source and destination IP address, and source and destination port information. The option contains alert message and packet which will be detected.

The simulated experiment is done with Snort [8, 9]. The system test uses common intrusion software to attack the protected system. The data sample is captured from this simulated environment. Though data analysis, the improved ID3 algorithm is proved to be effective.

In this experiment, Snort runs on Windows XP. It uses these softwares: Snort2.0, WinPcap_3_2_alpha1.exe, idscenter11rc4.zip, adodb495a, mysql-6.0.0, php-5.2.4-Win32, snort rules-snapshot-CURRENT, and acid-0.9.6b23.tar. The network speed is 100M/s.

Data from one or many ports is forward to a certain port by configuration Switch. Then the system can watch the network. SPAN can copy all communication from one Switch to Sniffer. Configure 6 one-way SPAN tasks: two input data stream monitoring, four output data stream monitoring. At first, capture all communication from some port, next, copy all communication to the target port.

B. Data Analysis

In the system environment above, Snort runs for a week consequently. Select the data record of Wednesdays′, divide the test data file by time, so the test file only contains certain intrusion attack. Because of Windows OS with vulnerabilities, DoS attack is chosen. The sample has 400 pieces of records. They meet the data analysis requirement basically. In the experiment records, redundant records were deleted. Detailed results can be searched in system log. The experiment data is processed by classification of individuals. After alert data is processed, we can get 4 alarm levels, 15 alert record [3]. In the sample, 0 means no intrusion, other levels increase by their number decrease.

TABLE I. DATA SAMPLE

RID	Attacks≥20	Alert Level	Target Status	Class
1	N	0	Safe	N
2	N	1	Attack	P
3	N	0	Safe	N
4	Y	2	Safe	P
5	N	1	Attack	P
6	N	3	Attack	N
7	Y	2	Safe	P
8	N	3	Safe	N
9	N	2	Attack	N
10	N	2	Attack	P
11	N	0	Safe	P
12	N	3	Attack	P
13	N	1	Attack	P
14	N	0	Safe	N
15	Y	3	Attack	P

The data sample is divided. At first, calculate information gain of each attribute. There are 9 class P and 6 class N. So, I (P, N) =0.970954. If selecting attribute attacks as classification attribute:

When attacks<20, I (6, 6) =1,

When attacks≥20, I (3, 0) =0.

Information entropy: E (attacks) =0.8.

Information gain:

Gain (attacks) =I (P, N)-E (attacks) = 0.170954.

Similarly,

gain(alert level)=0.2472991, gain(target status)=0.0784997. The attribute alert level has maximum information gain, it is used to classify the data sample. The sample is changed into four parts. Next, two non-leaf nodes are divided. At last, a decision tree is created. Fig. 2 shows it.

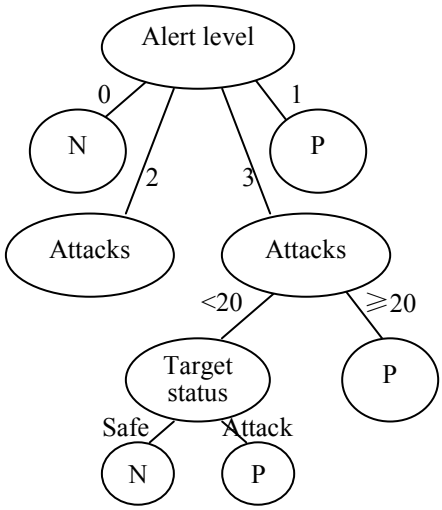

Figure 2. Decision Tree is Created by Original ID3 Algorithm

In the field, the experience of experts has shown that the attribute alert level isn′t important. There will be error in decision tree if alert attribute is root.

Selecting improved ID3 algorithm, new information gain values are calculated.

gain′ (attacks) =gain (attacks)*(15-9) =1.025724,

gain′ (alert) =gain (alert)*(15-11) =0.9891964,

gain′ (target) =gain (target)*(15-9) =0.470994.

Comparing ID3 algorithm and improved ID3 algorithm, table 2 is created.

TABLE II. COMPARISON OF INFORMATION GAIN BEFORE AND AFTER MODIFICATION

	Attacks	Alert Level	Target Status
Before modified	0.170954	0.247299	0.078499
After modified	1.025724	0.9891964	0.470994

From Table 2, the attribute attacks has maximum. So it is selected to divide data sample. The sample is divided into two parts. A decision tree with two non-leaf nodes is created. The tree is like figure 3.

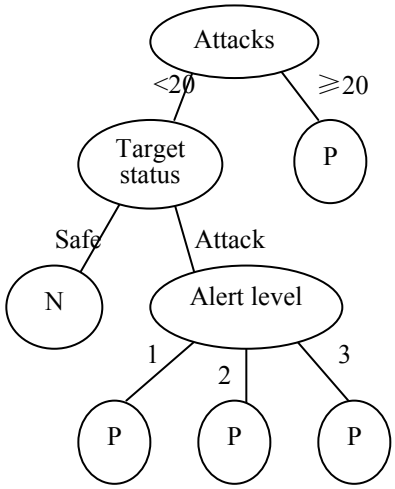

Figure 3. A Decision Tree is Created by Improved ID3 Algorithm

According to the traditional ID3 algorithm, alert level is the main property when classified. Level 1, level 2 and level 3 have different hazard to computer and communication. To attack with minimum level, it has hazard, too. Therefore, measures should be taken. This point can be reflected in improved ID3 algorithm. The data sample is divided by attribute attacks. So, the more the number of attacks, the more serious the hazard is. Once the alarm rings, measures should be done. To event with less attack, measures should be done, too. From the decision tree created by improved ID3 algorithm, the event with attacks<20, has hazard, too. We can get rules like these:

If attacks≥20 then alarm rings;

If ((attacks<20) and (target status=attack) and (alert level=1 or alert level=2 or alert level=3)) then alarm rings; and so on.

The rule base decreases after the ID3 algorithm is improved. Next, new rules are put into base for detecting new intrusion. The detection speed is elevation by improving the algorithm. Fig. 4 shows this.

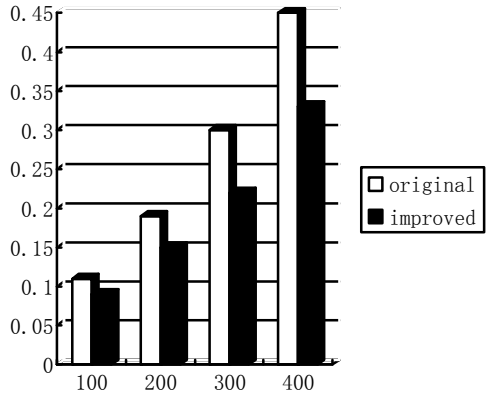

Figure 4. Comparison of Matching Time Before and After Modification (x: Records, y: Time)

V. CONCLUSION

Text heads organize the topics on a relational, hierarchical basis. For example, the improved algorithm is based on original ID3 algorithm, and maintains the classification accuracy of the original algorithm. The improved algorithm changes the way of selecting classification attribute of original ID3 algorithm. The decision tree created by improved algorithm has less nodes and paths. It means that there will be fewer rules, and the rules will be short and simple. Therefore, the detecting time of improved algorithm is less than original algorithm. At the same time, false alarm and omission alarm will be low.

REFERENCES

[1] Ming Fan, Xiaofeng Meng translated, Data mining techniques and concepts, Machinery Industry Press, Beijing, pp. 136-145, Feb., 2004.

[2] Wen Xu, Yang Zhang, "ID3 Algorithm Improvement On It", Computer Application and Digital Engineering, 2009, 10th, vol. 37: pp.19~21.

[3] Aidong Sun, Meijie Zhu and Shuqin Tu, "Improvement On ID3 Algorithm Based On attribute values", Computer Application Research, 2008, 12th, vol.29: pp.3011-3012.

[4] Mingqiu Song, Yun Fu, "Research Of Intrusion Detection Based On Protocol Analysis And Decision Tree", Computer Application Research , 2007, 12th, vol.24 : pp.171-172.

[5] Ming Huang, Wenying Niu and Xu Liang, "An improved decision tree classification algorithm based on ID3 and the application in score analysis", 2009 Chinese Control and Decision conference, pp.1876-1878

[6] Paul E.Utgoff, "Incremental Induction of decision trees", [EB/OL].http://www.cs.umass.edu/~utgoff/papers/mlj-id5r.pdf.[2008-2-01].

[7] Yonggui Zou, Chenhua Fan, "Improvement On ID3 Algorithm Based On the Importance Of Attribute Values", Computer Application, 2008, vol.28: pp.540-541

[8] http://www.snort.org

[9] Jinsong Song etal. Traslated, Snort 2.0 Intrusion Detection, National Defense Industry Press, Beijing, pp.73-104, Jun. 2004.