Application of ID3 Algorithm in Information Asset Identification

Hua Yong
Department of Foundation, the First Aeronautical College of Air Force, Xinyang, Henan, 464000, China
hy8106@163.com

Zhang Yunlong
Electronic Information Engineering College,Sias International University, Zhengzhou University, Xinzheng, Henan,451150,China
Email: zhang.yunlong317@163.com

***Abstract*—In the issue of information security risk evaluation, the asset, the threat and the vulnerability are the three most important elements. Information asset identification is a primary link of information security risk evaluation process. In this paper, the decision tree algorithm was applied into the identification of information assets; the basic process of ID3 algorithm are described; the data of information system property is classified by ID3 algorithm; the decision tree was made; and the rules are extracted for providing basis of information asset recognition.**

Key words: information Security; decision tree; ID3 algorithm

I. Introduction

Information security risk evaluation is the basis for an organization to ensure information security. Among the factors that information security risk evaluation involved, the information asset is the foremost important one. Information assets are the primary object of the information system security policy. The main goal of the asset identification is to find out which assets are most valuable for enterprises. Asset identification process can be divided into four steps: assets classification, assets information collection, assets object recognition and assets assignment. The process is actually a classification process. Asset identification is the scientific identification of the value of information assets, to determine the importance of it. The correctness and accuracy of the information asset identification for risk factor evaluation and comprehensive evaluation is essential.

The decision tree is a predictive modeling method for classification, clustering, and forecasting. As one of the core technology in data mining, decision tree has been widely used in many areas. This paper attempts to use decision tree ID3 algorithm to identify information assets, to create a decision tree with the greatest information gain ratio, and the extraction of the rules, and the data in this Information assets table is classified correctly.

II. Basic Principle of ID3 Algorithm

A Decision Tree

With the development of technology, computer, network, database technology is widely used in the daily management. All walks of life have accumulated a wealth of information and data. Database access and query operations cannot meet the requirements. People need to mine more important information from these massive data, such as the overall characteristics of the data description, to discover the interrelatedness of events and predict the development trend of things.

Most of the data mining methods use rule discovery or decision tree classification techniques to discover patterns and rules. Its core thought is some kind of induction algorithm. These methods usually mine the data in the database firstly, to generate rules and decision trees, then analyzing the new data and forecasting. The main advantages of these methods are readable by the rules and decision trees.

The decision tree classification algorithm in data mining is also an example-based inductive learning algorithm. It looks at a group of no order, no rules, examples, reasoning a decision tree that the formation of the classification rules. The algorithm uses a tree structure to represent a decision set, and using the classification of the data sample set to generate decision rules. Each non-leaf node of the tree represents an attribute test; its branches represent the test results. Each leaf node represents a category. In the decision tree-building process, you need to use pruning to cut the noise in the data and outsiders, thereby improving the reliability of the classification in the unknown data. There are two types of decision trees, classification trees and regression trees. The classification tree is for discrete variables, and regression trees are for continuous variables.

B ID3 Algorithm

In 1986, Quinlan proposed the famous ID3 algorithm. The ID3 algorithm is a decision tree classification algorithm based on information entropy. The core thought of the algorithm is to select properties on the decision tree nodes at all levels. The algorithm uses the highest information gain as the test attribute of the node, so the test for each non-leaf nodes can be up to classified information about the samples being tested. The sample set is divided into several sub-sets based on this property, and then let the entropy of the system go to minimum.

Tree generation algorithm (ID3):

Let S be the set of s data samples, assume that the decision attribute has m different values. C_i (i = 1, ..., m}) is the m different classes, s_i is the number of samples in the class C_i. Expectations of sample classification information can be

Funds Support: the science and technology research projects (102102210078) of Henan Science and Technology Department; the science and technology research projects (2010C520023) of Henan Education Department, and the scientific research project supported by SIAS international college, Zhengzhou University.

given by the following formula (1):

$$I(s_1,s_2,\ldots,s_m)=-\sum_{i=1}^{m}\frac{s_i}{s}\log_2(\frac{s_i}{s}) \qquad (1)$$

Attribute A has v different values $\{a_1, a_2, a_3, ..., a_v\}$, as the root of the decision tree, S is divided into v subsets $\{S_1, S_2, ..., S_v\}$, which S_j contains some samples of S, and they have the same value of a_j in A.

For a subset of S_j given formula (2):

$$I(s_{1j},s_{2j},\ldots,s_{mj})=-\sum_{i=1}^{m}\frac{s_{ij}}{s}\log_2(\frac{s_{ij}}{s}) \qquad (2)$$

The entropy of subsets divided by attribute A is given by equation (3):

$$H(A)=\sum_{j=1}^{v}\left[\left(\frac{S_{1j}+S_{2j}+\cdots+S_{mj}}{S}\right)I(S_{1j},S_{2j},\cdots,S_{mj})\right] \qquad (3)$$

The smaller the entropy value, the higher the purity of the Subsets. The information gain obtained by the branch from attribute A is given by equation (4):

$$\text{Gain}(A)=I(S_1,S_2,\ldots,S_m)-H(A) \qquad (4)$$

ID3 algorithm selected attribute A, the largest Gain(A), as the root of the sample set. Subsets of each branch use recursively the ID3 algorithm to build the decision tree nodes and branches, until the sample subsets belong to the same class. This approach makes the minimum average depth and the faster speed of the generated decision tree. A decision tree is generated.

C Decision Rules

Decision rules can be extracted from the decision tree. The method is: to create a classification rule from the root to the leaf nodes of each path, each attribute - value on the path is the rule antecedent (i.e., IF part) of a conjunctive item. Leaf node is after parts of the rules (i.e., THEN part). These rules can be used to classify new examples.

III. Application Examples

A Data Preprocessing

Information assets is a valuable corporate resource, it can exist in various forms of intangible, tangibles, hardware, software, codes, documentations, services, images et al. Confidentiality, integrity and availability are the three basic attributes in the evaluation of information security. The value of information assets in the risk evaluation is not only the economic value of the assets to measure, but mainly based on the reality and impact of confidentiality, integrity, availability of the information assets. Assets, which property safeties are different, have different values. Threats to assets, vulnerability and the security measures will impact on asset security attributes. Table 1 is an OA system risk evaluation data table, which assets consist of hardware assets, documents, assets and data assets and system assets. According to the importance and protection requirements of confidentiality, integrity and availability, with "5", "3" and "1" indicate the "high" level, "medium" level and "low" level, respectively.

Table 1. data table

serial number	Confidentiality	Integrity	Availability	Value
1	5	5	5	high
2	5	5	5	high
3	5	5	5	high
4	3	5	5	medium
5	3	5	5	medium
6	3	5	5	medium
7	1	1	1	low
8	1	1	1	low
9	5	5	1	medium
10	5	5	3	medium
11	1	5	5	low
12	1	5	5	low
13	5	3	1	low
14	5	3	1	low
15	5	3	1	low
16	1	3	1	low
17	1	3	1	low
18	5	3	1	low

B Calculating the Information Gain

1) To calculate the expectations required for sample classification. In Table 1, the attribute "value" has "high", "medium" and "low" three kinds of values, so the sample set will be divided into three categories: the ω_1 indicates "high", ω_2 indicates "medium", ω_3 indicates "low". By the data in table 1, $s_1=3$, $s_2=5$, and $s_3=10$. Take them into the formula (1):

$$I(\text{value})=I(s_1,s_2,s_3)=-\frac{3}{18}\log_2\frac{3}{18}-\frac{5}{18}\log_2\frac{5}{18}-\frac{10}{18}\log_2\frac{10}{18}$$
$$=0.9810$$

2) To calculate the information entropy of each attribute. For example, property "Confidential" has three values "5", "3" and "1", i.e., three sub-sets, respectively, to calculate the expectations of the three sub-sets:

Confidentiality="5": $s_{11}=3$, $s_{21}=2$, $s_{31}=4$

$$\text{I}(s_{11},s_{21},s_{31})=-\frac{3}{9}\log_2\frac{3}{9}-\frac{2}{9}\log_2\frac{2}{9}-\frac{4}{9}\log_2\frac{4}{9}$$
$$=1.0609$$

Confidentiality="3": $s_{12}=0$, $s_{22}=3$, $s_{32}=0$

$$I(s_{12},s_{22},s_{32})=0$$

Confidentiality = "1": $s_{13}=0$, $s_{23}=0$, $s_{33}=6$

$$I(s_{13},s_{23},s_{33})=0$$

$$H(\text{Confidentiality})=I(s_1,s_1,s_1)+\frac{3}{18}I(s_2,s_2,s_2)+\frac{6}{18}I(s_3,s_3,s_3)$$

$$=0.5304$$

$H(\text{integrity})=0.5721$, $E(\text{availability})=0.6554$

3) to calculate the information gain of attribute A. Obtained by the formula (3):

$$H(\text{confidentiality})=I(\text{value})-H(\text{confidentiality})$$
$$=0.9810-0.5304$$
$$=0.4506$$
$$H(\text{integrity})=I(\text{value})-E(\text{integrity})$$
$$=0.9810-0.5721=0.4089$$
$$H(\text{availability})=I(\text{value})-H(\text{availability})$$
$$=0.9810-0.6554$$
$$=0.3256$$

Comparing the above calculation results，information gain of "Confidentiality" is the largest, so it is chosen as a decision tree root. This property “Confidentiality" has three values, so have the three branches. Repeat the above steps, continue to split，then the decision tree will be generated. The results are shown in Figure 1.

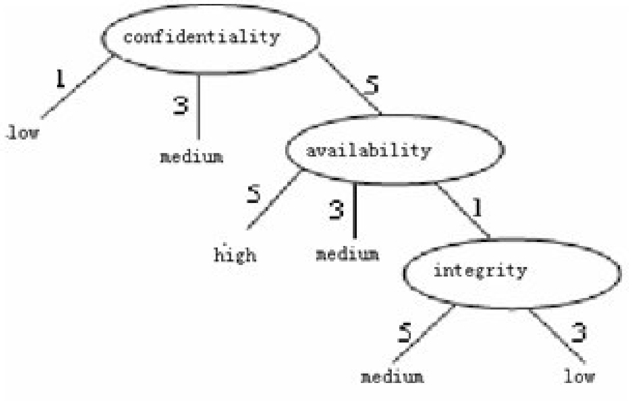

Figure 1. decision tree

C Extracting Classification Rules

The decision tree can be converted into the form of rules, in order to more clearly understand it, as follows:

IF "confidentiality"＝"1" THEN " value"＝"low";
IF "confidentiality"＝"3" THEN " value"＝"medium";
IF "confidentiality"＝"5" AND "availability"＝"5" THEN "value"＝" high";
IF "confidentiality"＝"5" AND "availability"＝"3" THEN "value"＝" medium";
IF "confidentiality"＝"5" AND "availability"＝"1" AND "integrity"＝"3" THEN "value"＝"low";
IF "confidentiality"＝"5" AND "availability"＝"1" AND "integrity"＝"5" THEN "value"＝"medium".

IV. Conclusions

The ID3 algorithm is a decision tree classification algorithm based on information entropy. The algorithm selects attributes with the largest amount of information gain as the test attribute of the current node. It makes the minimum amount of information needed by the data classification, and reflects the principle of minimum randomness. The ID3 algorithm is applied to the recognition of the value of information assets. Thus, we can get the value of the assets identification rules, and provide important support for information security risk evaluations.

References

[1] J. R. Quinlan. Induction of Decision Tree [J]. Machine Leaning, 1986(1): 81-106.

[2] J Mingers. An empirical comparison of pruning methods for decision tree induction[J]. Machine Leaning, 1989, 4(2):27-243.

[3] LI HE, WANG Shuyang. Research on the model and methods of information security evaluation. China Safety Science Journal, 2007, 17(2): 144-148.

[4] The International Organization for Standardization. Common criteria for information technology security evaluation [J].ISO/IEC15408,1999(E).

[5] ITSEC. Information technology security evaluation criteria 1.2[S]. Office for Offical Publications of the European Communities,1991.