Decision Tree Analysis on Environmental Factors of Insider Threats

Michihiro Yamada(✉), Koichi Niihara, and Hiroaki Kikuchi

Meiji University Graduate School, Meiji University,
Tokyo 164-8525, Japan
jam_1934@yahoo.co.jp

Abstract. In information security management, insider threat is one of the biggest threats. Since there are too many involved factors, it is not clear which factor plays the most significant role in malicious activities. Hence, this study aims to identify the factors of insider threat for security management viewpoint. We conduct an experiment from that a total of 198 subjects work to review sample web search engine and observed what they behaved. Our decision tree analysis reveals the typical characteristics of malicious ideas.

1 Introduction

In information security management, insider threat is one of the biggest threats. Since there are too many involved factors, it is not clear which factor plays the most significant role in malicious activities. Hence, this study aims to identify the factors of insider threat for security management viewpoint. Classifying behaviors into two classes, positive and negative, Hausawi conducted interviews with security experts [1]. This survey-based study is very useful for understanding insider's behaviors and collecting all possible features for malicious activities. However, survey and interview are not always true, e.g., subjects pretending to be honest and unintentionally protecting their organization. Moreover, it is not feasible to observe potential insider's every steps to perform malicious action.

In order to overcome the difficulties, in observation, we propose an experiment that subjects are employed to work to given task and observe the number of malicious activities of subjects.

We conduct an experiment from that a total of 198 subjects work to review sample web search engine and observed what they behaved. Our decision tree analysis reveals the typical characteristics of malicious ideas.

2 Experiment

In our experiment, we focus on an assignment of identities to users. If users share some common ID such as "administrator" with others, they tend to be malicious more often than users with individual IDs. Since it is impossible to figure out who makes misbehavior, the ID sharing user may think the malicious activities

C. Stephanidis (Ed.): HCII Posters 2017, Part II, CCIS 714, pp. 658–662, 2017.
DOI: 10.1007/978-3-319-58753-0_93

never be exposed. To verify how much malicious activities are increased when ID is shared, we divided a set of subjects into two groups; the first half assigned to a common ID, and the other half assigned to individual IDs.

The flow of experiment is as follows. First, all subjects login to a registration site by using the IDs of the crowdsourcing service. At that site, subjects are assigned the word list to be studied in a trial search service. Second, at the search site, subjects are divided into two groups; individual-IDs users and ID-sharing users groups. The individual-IDs users need to input their IDs of crowdsourcing site before login to the search site. While, the ID-sharing users are allowed to login without any information for access. At the search site, they tested the given 50 search words, and evaluate the quality of results as well as the performance of the search function.

If subjects complete the test with less than 50 words, we regard them as a malicious activity.

3 Experimental Results

3.1 Malicious Subjects

Table 1 shows the summary of experimental result. We show the numbers of malicious subject for their demographic attributes, e.g., sex, age, and affiliations.

Table 1. Maliious subjects with respects to demographic groups

	Shared IDs		Individual IDs		Total	
Group	Malicious	N	Malicious	N	Malicious	N
Sex male	13	51	11	58	24	109
Sex female	7	47	4	42	11	89
Age –19	1	1	0	0	1	1
Age 20–29	2	15	2	8	4	23
Age 30–39	9	35	4	41	13	76
Age 40–49	2	30	4	38	6	68
Age 50–59	2	12	2	10	4	22
Age 60–	4	5	3	3	7	8
Job office worker	5	22	5	26	10	48
Job public servant	1	1	0	0	1	1
Job self employed	7	28	3	29	10	57
Job parttime worker	1	9	0	10	1	19
Job houseworker	2	19	2	18	4	37
Job students	1	1	1	1	2	2
Job unemployment	1	9	3	12	4	21
Job others	2	9	1	4	3	13
Total	20	98	15	100	35	198

As a result, we observed that 20 ID-sharing users (out of 98) played malicious activity. The number of malicious users who shared a common ID is greater than that of individual-IDs users. Based on the experimental result, we analyze the set of malicious subjects in some methods, (1) Decision tree, and (2) Association rule mining, and (3) Logistic regression analyses.

3.2 Decision Tree Analysis

By a decision tree, node "Age" is chosen as the best classifier, which is at the root of tree, and plays a significant role for insider.

A decision tree reveals the logical conditions for determining a target attribute. Figure 1 shows the decision tree of malicious users, learned in R package "`rpart`". The target attribute is whether the subject is malicious or not. In this tree, nodes are logical conditions to classify subjects and the left branch means satisfied. By labels "Malicious/Honest", we denote the numbers of subjects in the node. For example, if user's age is over 55 (at the left sub tree of the root node) then 7 subjects are malicious except 1 honest (at `Sex = b`).

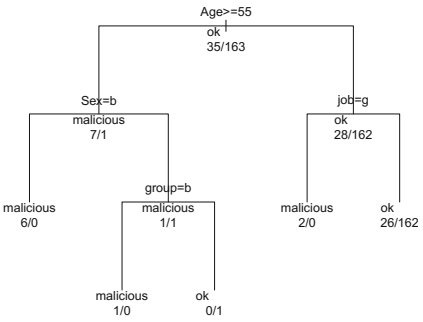

Fig. 1. Decision tree of malicious subjects

3.3 Association Rules Mining

To reveal the typical characteristic of insider related with combination of attributes, we extract some association rules by using R package, "arules". Table 2 shows the selected association rules. By Support and Confidence, we denote a joint probability $Pr(lhs, rhs)$, and a conditional probability $Pr(rhs \mid lhs)$, respectively. For example, No. 1 rule means that "If users use individual IDs and they are self-employed worker, then they are legitimate with 89% confidence. No. 5 means that "ID-sharing users sometimes (20% confidence) have played malicious activity".

The association rule shows "If individual-IDs users are 30's, they are legitimate".

3.4 Logistic Regression Analysis

A logistic regression is an analysis method to predict a conditional probability of event given conditions in a logistic model. We applied the logistic regression to the dataset of malicious subjects of some demographic attributes. Our model is of the form

$$\log \frac{Pr(malicious \mid x)}{1 - Pr(malicious \mid x)} = -1 - 0.05x_1 + 0.048x_2 + \cdots + 0.064x_{10}$$

where the coefficients of variables are given in Table 3.

Table 2. Assosiation rules

No	lhs	rhs	support	confidence	lhs.support	lift
1	{group = individual IDs, job = self-employed} =>	{Judge = ok}	0.1313131	0.8965517	0.1464646	1.089063
2	{group = individual IDs, Age = 40's} =>	{Judge = ok}	0.1717172	0.8947368	0.1919192	1.086858
3	{group = individual IDs, Age = 30's} =>	{Judge = ok}	0.1868687	0.902439	0.2070707	1.096214
4	{group = individual IDs, Sex=Male, job=self-employed} =>	{Judge = ok}	0.1111111	0.9166667	0.1212121	1.113497
5	{group = shard IDs} =>	{Judge = malicious}	0.1010101	0.2040816	0.4949495	1.154519

Table 3. Logistic regression analysis

	Estimate	Pr(>|t|)	Odds
(Intercept)	−0.107074	0.384287	2.41E-02
Group individual IDs	−5.42E-02	0.306387	6.78E-01
Sex male	0.048906	0.465707	1.41E + 00
Age	6.49E-03	0.023689 *	1.05E + 00
Job self-employed	0.031873	0.735564	1.38E + 00
Job office worker	0.097586	0.297715	2.18E + 00
Job other	0.087399	0.476033	1.86E + 00
Job part-time worker	−0.06025	0.566693	4.41E-01
Job public servant	0.668873	0.082308	2.90E + 07
Job student	1.012411	0.000336 ***	3.37E + 08
Job unemployment	0.06497	0.558746	1.74E + 00

4 Conclusions

We studied the factor analysis of malicious insider in total of 198 subjects with some conditions. Our experiment showed that sharing ID and Password could increase a risk of malicious insider by $\frac{1}{0.68}$ times than without sharing.

References

1. Hausawi, Y.M.: Current trend of end-users' behaviors towards security mechanisms. In: Tryfonas, T. (ed.) HAS 2016. LNCS, vol. 9750, pp. 140–151. Springer, Cham (2016). doi:10.1007/978-3-319-39381-0_13