World Applied Sciences Journal 29 (12): 1611-1614, 2014
ISSN 1818-4952

DOI: 10.5829/idosi.wasj.2014.29.12.13988

Detection of an Intruder in an Information System Through the Check of Statistic Hypotheses by Certain Statistic Criteria

Jurij Aleksandrovich Gorbunov, Lev Nikolaevich Krotov and Elena L'vovna Krotova

Perm National Research Polytechnic University,
Komsomolsky Ave. 29, 614990 Perm, Russian Federation

Abstract: The article deals with the problem of detecting an intruder in an information system in its function as an issue of checking the statistic hypothesis of homogeneity of data extractions. The solution of the problem based on the application of the statistical analysis to the results of monitoring the user activity in the system is suggested. Two methods were considered in detail based on statistic criteria of comparison of two average parent populations, the variances of which are known for large independent extractions and comparison of two average standard parent populations, the variances of which are unknown and equal. The conclusion is made on the possibility of application of these methods for solution of this problem within the intrusion detection systems.

Key words: Intrusion detection systems · Statistical methods · Homogeneity criteria

INTRODUCTION

Presently, we can surely say that information systems are applied in any spheres of social life. These systems have been integrated so deeply into the existing processes that the issue of information security becomes especially important with respect to information processed by them, namely the issue of confidentiality, integrity and accessibility of the information.

Timely detection of unauthorized intrusion into the system and detection of the intruder play an important role in ensuring the security of information systems. The research related to solution of the problem of detection of an information system intrusion has been being actively carried out during the last two decades. Some of them consider the methodological part of the problem, the issues of taxonomy of the intrusion detection systems [1]. A separate and quite large part of the research is dedicated to detection of intrusions into the network routing protocols, such as BGP (Border Gateway Protocol) [2, 3]. The carried out researches suggest applying the intrusion detection systems based on the hidden Markov models allowing to detect attacks, which use XSS and SQL injections [4]. During the recent years, another approach to the solution of the problem of detection of an information system intrusion has formed, which is based on application of statistical methods of intruder detection [5-9].

Within this article, we considered the methods of an assumed intruder detection based on the statistical analysis of information about his activity in the system and activity of legal users. The methods are based on formulation of respective statistic hypotheses and their testing using statistic odd tests.

Methodology: Let us assume that there are two independent extractions $X = (X_1,...,X_n)$ and $Y = (Y_1, ...,Y_n)$, which describe the same process but were received at different time or under different circumstances. It is required to define whether these are extractions from the same distribution or the observations distribution law changed between the extractions. In other words, within the context of this research, the question is whether these data are related to the behavior of the same user and whether the user is an intruder or not.

Generally, the problem can be formulated in the following manner:

Corresponding Author: Gorbunov, Perm National Research Polytechnic University, Komsomolsky Ave. 29, 614990 Perm, Russian Federation.

62	76	60	64	61	95	79	91	89	90	76	90	62	98	61	71	96	85	94	74	85	95	83	67	73	82	68	81	91	93	68	62
93	64	62	68	83	96	89	75	88	62	62	99	66	69	64	92	89	60	78	68	71	68	96	83	96	71	90	67	87	87	61	92
68	89	73	74	68	80	73	88	74	64	76	80	80	96	64	62	81	92	63	87	83	93	63	80	81	64	92	72	93	99	93	96
88	60	88	94	88	93	61	78	91	80	95	63	99	71	87	98	76	73	79	77	77	93	73	69	95	72	82	72	96	87	92	64
97	73	75	62	67	68	73	94	83	74	85	86	71	78	94	67	90	95	76	97	83	79	75	76	63	74	62	93	81	87	69	92
76	77	84	89	83	69	76	73	71	82	74	66	64	85	78	62	75	79	89	80	92	97	84	60	62	96	85	94	80	68	96	92
94	77	83	73	61	83	96	84	90	92	61	74	91	69	68	95	68	99	98	87	97	99	75	99	89	76	95	72	94	95	74	89
75	86	83	93	93	74	89	97	82	61	61	98	99	62	76	60	79	73	84	97	91	94	99	70	78	68	89	86	84	68	84	91
90	64	72	92	77	87	68	63	77	81	86	71	98	73	79	63	89	87	79	84	61	96	77	92	91	74	79	65	99	67	65	91
62	78	89	65	80	73	98	70	68	62	74	70	88	80	85	92	84	87	91	83	67	93	95	74	78	87	79	97	86	68	79	61
69	75	72	77	94	75	78	95	97	73	98	80	69	96	61	62	72	81	89	92	70	69	89	78	63	98	98	92	93	86	82	84
68	81	82	99	75	94	92	72	79	78	69	99	75	78	99	99	73	86	66	70	64	73	88	82	76	80	65	95	85	64	68	86

Fig. 1: The source data of a legal user.

Let us assume that $X = (X_1, ..., X_n)$ is an extraction from the distribution ***Z([xi])*** with a certain unknown distribution function $F_1(x)$ and $Y = (Y_1, ..., Y_n)$ is the extraction from the distribution ***Z([eta])*** with the unknown distribution function $F_2(x)$. The task is to test the *homogeneity hypothesis* H_0: $F_1(x)=F_2(x)$.

12 randomly generated data arrays with the length of 32 values $\boldsymbol{L_1..L_{12}}$ were taken as the source data. The values in the form of an array are represented in *Figure 1* and correspond to the conditional sequence of the values of event logging time.

We take an array of 32 values $\boldsymbol{K_1}$ as the source data related to the intruder's behavior:

[109, 86, 75, 86, 92, 103, 90, 78, 110, 73, 88, 105, 75, 105, 105, 79, 109, 71, 92, 85, 103, 97, 80, 72, 74, 108, 80, 106, 80, 73, 96, 91]

and the array of 15 values $\boldsymbol{K_2}$:

[109, 86, 75, 86, 92, 103, 90, 78, 110, 73, 88, 105, 75, 105, 105],

as we assume a short time will be required to detect the intruder after he starts his activity in the system.

Body of the Work: It is suggested to use the criterion of comparison of two average parent populations, the variances of which are known for large independent extractions, as one of the applied criteria [10].

As the arrays of the $\boldsymbol{L_1..L_{12}}$ values correspond to the conditional sequence of the values of the event logging time, we can calculate the array of arithmetic averages for each time point – ***L***. The found values array describes the activity of a legal user.

Using *n* and *m,* let us denote the volumes of large *(n* > 30, m > 30) independent extractions, by which we can find the respective average $\bar{l}$ and $\bar{k}_1$ as well as the population variances *D(L)* and $D(K_1)$:

80.24739583 *D(L)= 10.64627436*
89.875 $D(K_1)$= *174.3064517*

We find the required significance level [alpha] and test the null hypothesis H_i: $M(L) = M(K_1)$ about the equality of mathematical expectations (the population means) of two normal parent populations with known variances (in the case of large extractions) at the competing hypothesis H_1: $M(L) \neq M(K_1)$.

To do it, we need to calculate the observed value of the criterion

$$Z_{obs} = \frac{\bar{l} - \bar{k_1}}{\sqrt{\frac{D(L)}{n} + \frac{D(K_1)}{m}}}$$

and find the critical point z_{cr} using the table of the Laplace's function from the equation:

$$\Phi(Z_{cr}) = \frac{1-a}{2}$$

where **[alpha]** is the significance level.

If $|Z_{ob}| < z_{cr}$, the null hypothesis is not rejected, both data arrays ***L*** and $\boldsymbol{K_1}$ characterize a legal user's activity.

And if $|Z_{obs}| > z_{cr}$, the null hypothesis is rejected, which evidences the breach of the system security.

For the source data provided above and the significance level **[alpha]** =0.05, we have:

Z_{obs}= *4.004634155* z_{cr} = *1.96*

As $|Z_{obs}|$ is much larger than z_{cr}, the sample values differ considerably. The *k1* data evidence the breach of the system security.

In practice, the quickness of detection of system intrusion is also equally important. In this case, for small independent extractions, we can use the criterion of comparison of two average normal parent populations, the variances of which are unknown and equal [10].

In this case, there are less data about the assumed intruder, namely 15 values grouped into the array $\boldsymbol{K_2}$. Let us assume that we do not know much about the activity of the legal user; then we extract the first 20 values of the $\boldsymbol{L_1}$ array in the $\boldsymbol{L_{min}}$ array

[62, 76, 60, 64, 61, 95, 79, 91, 89, 90, 76, 90, 62, 98, 61, 71, 96, 85, 94, 74],

which we will use in this method.

Using these small extractions ($n < 30$, $m < 30$), we find the respective sample averages $\overline{l_{\min}}$ and $\overline{k_2}$ and the corrected sample variances s^2_{Lmin} and s^2_{K2}.

Let us assume that the population variances are equal but unknown.

In order to test the null hypothesis H_i at the preset significance level **[alpha]**: $M(L_{min}) = M(K_2)$ about the equality of mathematical expectations (population means) of two normal parent populations with unknown but equal variances (in the case of small independent extractions) at the competing hypothesis H_1: $M(L_{min}) \neq M(K_2)$, it is necessary to calculate the observed value of the criterion

$$T_{obs} = \frac{\overline{l_{\min}} - \overline{k_2}}{\sqrt{(n-1)s^2_{L\min} + (n-1)s^2_{k2}}} \cdot \sqrt{\frac{n.m.(n+m-2)}{n+m}}$$

Then, using the table of critical distribution points of Student, the preset significance level [alpha] and the number of degrees of freedom $p=n+m-2$, *we find* the critical point $t_{two\text{-}way\ cr}([alpha], p)$.

If $|T_{obs}| < t_{two\text{-}way\ cr}([alpha], p)$, the null hypothesis is not rejected, both data arrays $\boldsymbol{L}$ and $\boldsymbol{K_1}$ characterize a legal user's activity.

And if $|T_{obs}| > t_{two\text{-}way\ cr}([alpha], p)$, the null hypothesis is rejected, which evidences the breach of the system security.

However, this case has some restrictions. If the corrected variances s^2_{Lmin} and s^2_{K2} differ considerably, it is necessary to test the hypothesis of equality of the parent variances using the criterion of Fisher & Snedecor first. [10].

To do it, we need to find F_{obs} – the ratio of the larger variance to the smaller one. If the s^2_{Lmin} variance is considerably greater than the s^2_{K2} variance, we take the hypothesis H_1: $D(X) > D(Y)$ to be the competing hypothesis. If the s^2_{Lmin} variance is considerably smaller than the s^2_{K2} variance, we take the hypothesis H_1: $D(X) < D(Y)$ to be the competing hypothesis.

Using the Fisher & Snedecor's table of critical distribution points F, we find the critical point $F_{cr}([alpha], p_1, p_2)$, by the significance level [alpha] and values of the degrees of freedom $p_1=n-1$ and $p_2=m-1$.

If $F_{obs} < F_{cr}([alpha], p_1, p_2)$, there are no reasons to reject the null hypothesis about the equality of parent variances. The assumption on the equality of the parent variances is confirmed and this method is applicable for solution of the assigned problem.

If $F_{obs} > F_{cr}([alpha], p_1, p_2)$, the null hypothesis about the equality of parent variances is rejected and this method is not applicable for solution of the assigned problem.

For the source data provided above, we have:

$\overline{l_{\min}} = 78.7$ $\quad$ $\overline{k_2} = 92$

$s^2_{Lmin} = 178.51$ $\quad$ $s^2_{K2} = 164.5333333$

The corrected variances differ slightly, the parent variances are assumed comparable and this method can be applied.

$T_{obs} = 2.964028521$

For the level of significance [alpha]=0.05 and the number of the freedom degrees $p=33$ $(n=20, m=15)$

$t_{two\text{-}way\ cr}(0.05;33) = 2.04$

The observed value of the T_{obs} criterion is more than the critical value t_{cr} and the sample values differ considerably. According to the accepted axiom and the results of comparison, the second data were correctly identified as the data describing the behavior of an intruder.

Summary: The methods suggested above were tested on several dozens of source data sets. The results prove the applicability of such methods for solution of the problem of detecting an intruder in an information system.

Further, we plan to consider the possibility to apply the respective weight coefficient characterizing the probability of each value of source data about the activity of a legal user, which will make it possible to use large data volumes for building a profile and to determine the restrictions of the suggested methods' applicability.

REFERENCES

1. Almgren, M., 2003. Consolidation and Evaluation of IDS Taxonomies. In the Proceedings of the Eighth Nordic Workshop on Secure IT Systems, Nord Sec., pp: 57-70.
2. Lad, M. *et al*., 2006. PHAS: A Prefix Hijack Alert System. In the Proceedings of the 15th USENIX Security Symposium, pp: 153-166.
3. Huston, G., M. Rossi and G. Armitage, 2011. Securing BGP - A Literature Survey. IEEE Communications Surveys and Tutorials, 2: 199-222.
4. Ariu, D., R. Tronci and G. Giacinto, 2011. HMMPayl: An Intrusion Detection System Based on Hidden Markov Models. Computers & Security, 30(4): 221-241.
5. Gorbunov, Y.A., 2009. Using the Stochastic Approach to Distinguish the Legal User's Profile from the Profile of an Intruder. Eds., Gorbunov, Y.A., I.A. Zhuykov and E.L. Krotova, Review of Applied and Industrial Mathematics, 16(3): 460-461.
6. Gorbunov, Y.A., 2010. Determination of the Parameters of the Mathematic Model of a Workstation User Profile. Eds., Gorbunov, Y.A., I.A. Zhuykov and E.L. Krotova, Review of Applied and Industrial Mathematics, 17(1): 109.
7. Nesterenko, V.A., 2006. Statistical Methods of Detection of Security Breaches within a Network. Information Processes, 6(3): 208-217.
8. Karaychev, G.V. and V.A. Nesterneko, 2010. Detection of Abnormal Activity in a Network by Statistical Analysis of IP Packet Headers. News of Higher Educational Institutions. The North-Caucasian Region. Natural Sciences, 4: 13-17.
9. Tumoian, E. and M. Anikeev, 2005. Network-based Detection of Passive Covert Channels in TCP/IP. In the Proceedings of the 2005 IEEE Conference on Local Computer Networks: LCN '05, Washington, DC, pp: 802-809.
10. Gmurman, V.E., 2004. Guidance for Solution of Problems in the Theory of Probabilities and Mathematical Statistics: Guide for Tertiary Students. Moscow: Vysshaya Shkola, pp: 404.