DOI 10.29042/2018-4668-4673

Helix Vol. 8(6): 4668- 4673

Insiders Detection in Computer Systems Based on Datamining Technique

[1]Ilyas I. Ismagilov, [2]Linar A. Molotov, [3]Igor V. Anikin, [4]Alexey S. Katasev, [5]Dina V. Kataseva

[1,2] Kazan Federal University

[3,4,5] Kazan National Research Technical University named after A.N. Tupolev

Email: molotov.linar@mail.ru

Received: 02nd November 2018, Accepted: 28th November 2018, Published: 31st December 2018

Abstract

Insider threats are among the most frequent and dangerous for modern organizations. Therefore, the problem of timely detection of internal information security incidents is relevant. Due to the fact that many of the approaches to protection against insider attacks are not effective enough, recently an approach based on identifying anomalous user behaviour has often been used to solve the problem of identifying internal incidents. The use of this approach involves the construction of a reference profile of user behaviour and the further identification of facts of "atypical" behaviour. This work is devoted to solving the actual problem of detecting events in computer systems. These events occur when the access control rules are violated when an attacker tries to enter the system, posing as another user. Data mining technology is used to build profiles of the user's reference behaviour and to identify anomalies. Hidden patterns in user behaviour are revealed using the decision tree construction algorithms C4.5, search for associative rules FPGrowth, sequential data analysis. The types of internal incidents and ways to determine their criticality are determined. The developed technology is implemented in the tool complex of programs. It was tested on the example of solving the problem of detecting anomalies in the behaviour of users working with programs running the Windows operating system. In this case, the attacker did not try to hide his actions when working in the system. Identified system errors when recognizing an intruder. The error of the first kind was 20%, and the error of the second kind - 10%.

Keywords

Information Security, Internal Intruders, Data Mining, Anomaly Detection

Introduction

Currently, the main trend of global economic development is the development of the digital economy. The basic foundations of this economy are e-commerce and e-business. In this connection, the urgency of the task of developing information systems of organizations on the basis of digital economy technologies and their integration with e-commerce systems is increasing [1,2].

The emergence of the digital economy is inextricably linked with the development of risks in various subject areas, in particular, in the tasks of ensuring information security (IS). In this case, an urgent task is the timely detection and investigation of incidents related to the violation of the security of a computer system (CS). This is due to the fact that organizations often experience significant problems due to IS incidents, mainly related to disruptions in the functioning of the IT infrastructure, reputational costs and financial losses.

Today, the greatest relevance is acquired by the task of detecting violations of information security policies by legal users who have full rights to access a computer system in an organization. Such users are called insiders.

In this case, the total number of threats associated with the actions of insiders is three times greater than the number of threats from external users who are not employees of the companies in which these threats are realized.

On the other hand, existing approaches to protecting against them are often not sufficiently effective. Many internal incidents remain outside the scope of the remedies used, in particular, unauthorized actions taken under the account of a legal user for whom the relevant actions are allowed.

In order to detect such violations, special systems for detecting anomalies are of particular interest. The work of such systems is carried out in two stages. First, the construction of reference profiles of typical user behaviour is performed. Then, in the active phase of the system operation, a comparison of the current user's behaviour profile with its standard is made. If the deviation threshold is exceeded, it is possible to notify the administrator about the suspicious behaviour of the user. Facts are revealed when the user begins to behave "atypically." This method of detecting anomalous user behaviour is superior in efficiency to signature analysis methods. This approach has found wide application in intelligent systems for detecting attacks in computer networks [3,4], as well as in systems for detecting anomalous user behaviour [5-12].

The main element of anomaly detection systems is the knowledge base, which contains reference profiles of typical user behaviour. Consequently, the problem arises of the formation of these profiles. To effectively solve this problem, it is important to use Data Mining algorithms [13-17], which allow you to extract knowledge from the accumulated data. Realization of such algorithms is dedicated to the present work.

Methods

To implement the initial stage of operation of any anomaly detection system, it is required to form a standard of a legal user. In relation to the problem being solved, this standard should be formed on the basis of the intellectual analysis of

 © 2018 The Author (s); Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

the user's behaviour during his work with the programs installed on the computer. The source of the source data can be the Security EventLog, which is built into Windows.
Consider the structure of the developed system for detecting anomalous user behaviour, presented in Figure 1.

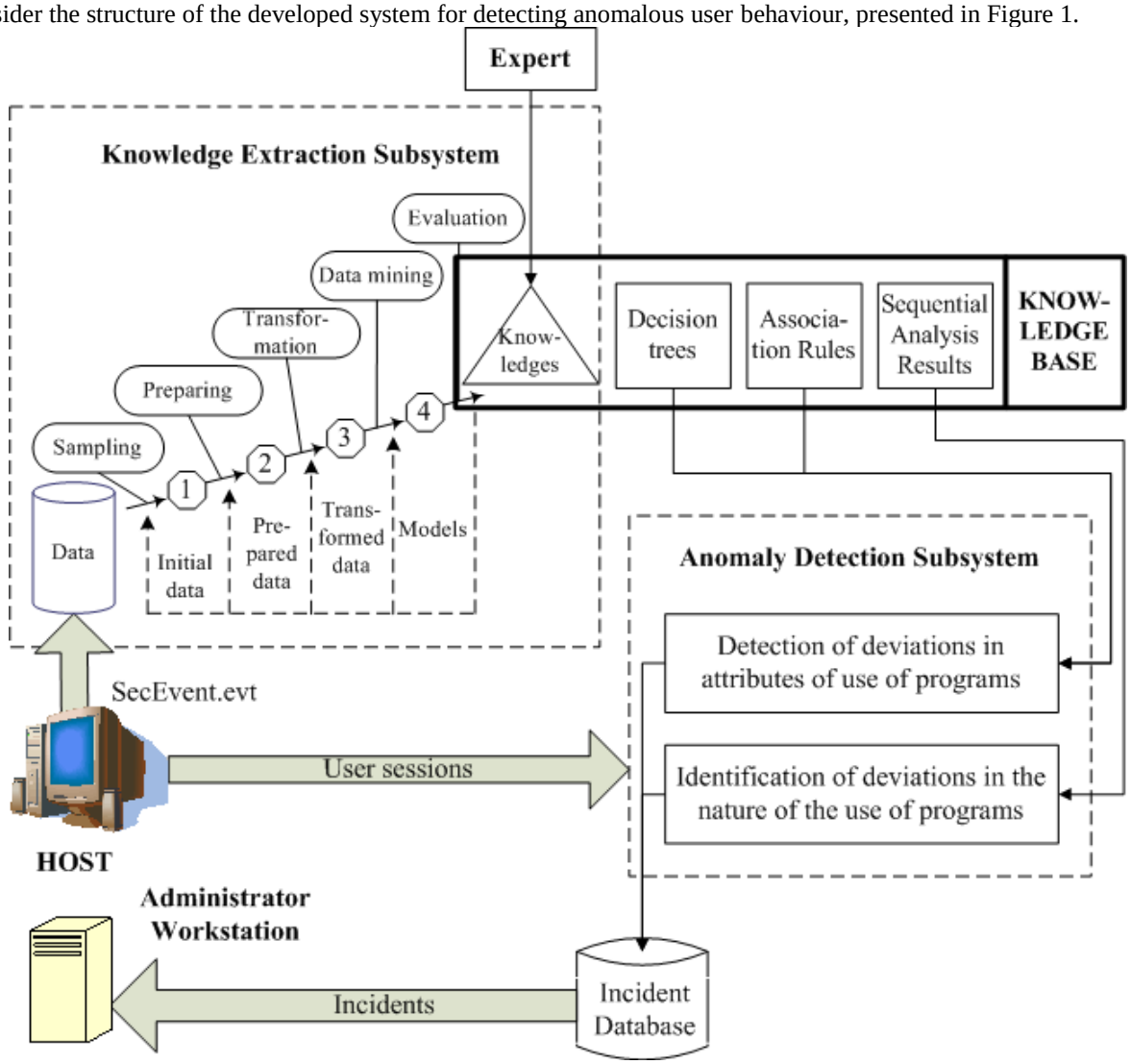

Figure 1: Structure of the System for Detecting Abnormal User Behavior

As can be seen from the figure, the developed system includes the following subsystems:
1) learning subsystem, in which the formation of reference profiles of typical user behavior is made, which constitute the knowledge base of an intelligent system (herewith, decision trees [18,19], associative rules [19,20] and sequential analysis methods [21.22]);
2) anomaly detection subsystem, which is responsible for applying the generated knowledge base to detect anomalous user behavior.
The abnormalities associated with the launch of executable programs at an unspecified time, as well as anomalies associated with the violation of the application launch sequence, are considered as atypical user behavior. All incidents are matched with criticality weights, and this information is entered into the database.
The training phase of the system is associated with the construction of reference profiles of typical user behavior. Let $S = \{s_1,...,s_q\}$ - all entries from the *Security EventLog*. Moreover, any record can be specified in the form $\langle d_i, t_i, type_i \rangle$, where $type_i$ is the type of event, d_i, t_i - are the date and time of this event.
The work of the knowledge detection subsystem consists of a series of successive stages.
1. Sampling the initial data. At this stage, a subset of records is formed $\overline{S} \subseteq S$, $\overline{S} = \{s_1,...,s_n\}$, which refer to the moments of the start or end of user programs.
2. Pre-processing of initial data, consisting in their cleaning, deletion of redundant data, deletion of irrelevant by the time events, as well as the concatenation of repetitive events.

 © 2018 The Author (s); Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

3. Data transformation, which consists in converting records from the set $\overline{S}$ to the required form from the point of view of the selected analysis algorithm. In particular, the parameters d_i and t_i must be transformed into a discrete representation of the type $\aleph_{date}$ and $\aleph_{time}$, where $\aleph_{date}$ = the beginning of the week, the middle of the week, the end of the week, the weekend, and $\aleph_{time}$ = early morning, morning, middle of the day, end of the day, evening, night, deep night.

4. Data Mining - the stage of direct application of data analysis algorithms. Consider the features of the selected algorithms in more detail.

Building a decision tree.

In this paper, when constructing a decision tree, the classical algorithm C4.5 [18] is implemented. The constructed tree is a classification rule that allows you to determine the day of the week and the time when the user started certain programs on the computer.

Search for associative rules.

The association rule search is based on transaction analysis. We denote $s_i \in \overline{S}$ as the elements of transactions, and each of the transactions can be specified in the form $\langle d_i, t_i, type_i \rangle$, where $d_i \in \aleph_{date}, t_i \in \aleph_{time}, type_i \in \aleph_{type}$. In contrast to the decision tree model, when searching for associative rules, not a complete system of rules is formed, but a small number of them. In addition, each rule is written in the following form:

$$X \to Y, C, I, \quad (1)$$

where X and Y are the elements of the corresponding sets $\aleph_{date}$, $\aleph_{time}$, $\aleph_{type}$ that appear in the analyzed transactions together; C - accuracy of the generated rules; I - improvement (is the obtained rule more useful than random guessing of the class of solutions).

In this work, the *FPGrowth* algorithm was used to search for associative rules. The table shows examples of found association rules.

Association Rule	Rule Interpretation
time = < Early morning >→ *date* = < Midweek >	If the user's working day starts in the early morning, then most likely it is the middle of the week.
type = < System sign-on >, *date* = < Beginning of the week >→ *time* = < Morning >	At the beginning of the week, the user starts work in the morning.
type = <*Nero* >→ *date* = < End of the week >	Nero is launched at the end of the week.

Table 1: Examples of Found Association Rules

Sequential analysis.

Sequential analysis is similar to the search for associative rules. However, the following user sessions are used as transactions*:*

$$Session = \{type_1, ..., type_n\}. \quad (2)$$

It can be seen that each session consists of a sequence of programs launched by the user on the computer.

The analysis of transactions of the form (2) is performed in order to search for associative rules and highlight in them sets of sequentially launched programs. To search for such sets of programs, an algorithm for constructing frequency trees has been implemented.

5. Assessment of the knowledge gained.

The stage of searching for knowledge ends with their assessment. For this purpose, an expert is involved who interprets the knowledge gained, performs their analysis and gives his own assessment. Assessment of the knowledge gained is to determine their adequacy and the possibility of practical use in the subsystem of detecting anomalies.

Results and Discussion

Consider the operation of the system in identifying abnormal user activity. At this stage, an analysis is made of the events recorded in the Security EventLog log, a further comparison of the user's behaviour with the reference one and the identification of anomalous activity. In this case, the following categories of incidents are distinguished:

1. The timing of the launch of the program p differs from the previously constructed rules of the decision tree. For such incidents, the distances $\rho(p_s)$ (from 1 to 5) to the closest leaves of the decision tree marked by the p . program are calculated on the ordinal scale. The severity of such incidents is determined by the formula

 © 2018 The Author (s); Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

$critical(i_j) = critical(p) \circ \min_s \rho(p_s)$, where $critical(p) \in \{1,...,5\}$ is the criticality level of the program under study p, which is determined by the expert.

2. The associative rule $A \to B$ is broken when an event A occurs. The criticality of such incidents is determined by the formula $critical(i_j) = critical(A \to B) \cdot C$, where $critical(A \to B)$ is the level of criticality of the rule, which is determined by the expert, and $C \in [0;1]$ determines the accuracy of the rule being investigated.

3. The software launch sequence is not executed in *k*-sequences. Criticality of such incidents $critical(i_j) = critical(l_k) \cdot \min_k \frac{Support(X \cup Y_k')}{Support(X \cup Y_k)}$, where is the *k*-sequence of the generated knowledge base, at the beginning of which the event is located; X ; Y_k' - the expected continuation of the *k*-sequence, and Y_k - the real continuation of the *k*-sequence.

Let $I_A = \{i_1,...,i_k\}$ be a number of security incidents related to an account u over a period of time T. Then the total criticality of incidents $C(u)$ related to the user account u is determined according to (3):

$$C(u) = \sum_{t=1}^{k} critical(i_t). \quad (3)$$

When performing a $C(u)$ value analysis, a decision is made about the presence of abnormal user behaviour. For an IB administrator, this level is represented in color by determining the *CI* value (color incident) as follows:

$$CI = \begin{cases} green, if\ C(u) \le C^{yellow}(u) \\ yellow, if\ C^{yellow}(u) < C(u) < C^{red}(u) \\ red, if\ C(u) \ge C^{red}(u) \end{cases}$$

The formation of threshold levels $C^{yellow}(u)$ and $C^{red}(u)$ is as follows. Let $U = \{u_i\}_{i=1}^{N_{user}}$ be user accounts, and $C(u_i) \ge 0$ - the level of incidents emanating from them, determined according to (3) when training the system. The formation of reference profiles of user behaviour occurs in the absence of account compromise. In this regard, the $C(u_i)$ value is valid. Let $K(u_i)$ be the set of knowledge units (decision trees, association rules, sets of jointly launched programs) for u_i obtained during the system training and included in the user behaviour profile.

At the end of the training system is setting up its parameters by forming levels $C^{yellow}(u)$ and $C^{red}(u)$ for accounts $u_i \in U$. At the setup stage, statistics on users' work in the system in the test mode is collected, and the time for collecting test data should be comparable to the time of training the system.

Let $E(u_j, K(u_i))$ be the distance between the behaviour profile of the i-th user and the behaviour of the j-th user at the testing stage. Collecting information about the u_i user at various points in time $0 < T_i \le T_{test}$ during testing determines and normalizes the distribution of his distances $P_i^t(r)$ - from the reference profile of u_i user behaviour, and $P_i^f(r)$ - from the reference profiles of the behaviour of other users. Then, when anomalies are detected, the $P_i^t(r)$ value determines the probability that the criticality of the incident is green (not critical), and the probability that the criticality of the incident is red (critical).

Then $C^{yellow}(u)$ and $C^{red}(u)$ are defined as follows:

$$C^{yellow} = r \text{ for which } \frac{P_i^t(r)}{P_i^f(r)} = 2,\ C^{red} = r \text{ for which. } \frac{P_i^t(r)}{P_i^f(r)} = \frac{1}{2} \quad (4)$$

For the IB administrator, the hue and color of the IB incidents is calculated according to expression (5):

$Color = \left(\overline{P_i^f(r)} \cdot 255, \overline{P_i^t(r)} \cdot 255, 0\right)$, where

$$\overline{P_i^t(r)} = P_i^t(r) \cdot k,\ \overline{P_i^f(r)} = P_i^f(r) \cdot k, \text{ where } k = \max\left\{\frac{1}{P_i^t(r)}, \frac{1}{P_i^f(r)}\right\}, \quad (5)$$

where $\overline{P_i^f(r)} \cdot 255$ – is the saturation of red, $\overline{P_i^t(r)} \cdot 255$ – is the saturation of green.

 © 2018 The Author (s); Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

Summary
The proposed approach of identifying insiders is implemented in a software product used to detect abnormal user activity in computer systems running Windows. At the same time, violating users did not try to disguise their actions as legal. The experiment showed the presence of a relative amount of classification errors of the 1st kind at the level of 0.2 (20%), and a relative amount of errors of the classification of the 2nd kind of 0.1 (10%), which is acceptable for solving the problem.

Conclusion
The proposed approach makes it possible to identify particularly critical information security incidents in computer systems that are associated with the compromise of a legal user account. The application of the obtained results in practice can significantly increase the degree of protection of computer systems from internal violators.

Acknowledgements
The work is performed according to the Russian Government Program of Competitive Growth of Kazan Federal University. This work was supported by the Russian Federation Ministry of Education and Science, project № 8.6141.2017/8.9.

References
[1] Ismagilov I.I., Belov A.I. Methodological aspects of choosing a portfolio of projects on integration of corporate information systems with e-commerce tools // Kazan Economic Vestnik. – 2010. – Vol.21, Is.4. – P. 64-69.
[2] Ismagilov Ilyas I., Khasanova Svetlana F., Zinov'ev Pavel A., Complex engineering systems: rational choice of evolutionary projects // REVISTA PUBLICANDO. - 2018. - Vol.5, Is.16. - P.409-420.
[3] Ismagilov I.I., Khasanova S.F., Katasev A.S., Kataseva D.V. Neural network method of dynamic biometrics for detecting the substitution of computer // Journal of Advanced Research in Dynamical and Control Systems. V. 10. – P. 1723-1728.
[4] Katasev A.S., Kataseva D.V. Neural network diagnosis of anomalous network activity in telecommunication systems // Proceedings of IEEE Conference Dynamics of Systems, Mechanisms and Machines, Dynamics 2016.
[5] Chattopadhyay P., Wang L., Tan Y.-P. Scenario-based insider threat detection from cyber activities // IEEE Transactions on Computational Social Systems 5(3). – P. 660-675
[6] Le D.C., Zincir-Heywood A.N. Evaluating insider threat detection workflow using supervised and unsupervised learning // Proceedings of 2018 IEEE Symposium on Security and Privacy Workshops, SPW 2018. – P. 270-275.
[7] Meng F., Lou F., Fu Y., Tian Z. Deep learning based attribute classification insider threat detection for data security // Proceedings of 2018 IEEE 3rd International Conference on Data Science in Cyberspace, DSC 2018. – P. 576-581.
[8] Dahmane M., Foucher S. Combating insider threats by user profiling from activity logging data // Proceedings of 2018 1st International Conference on Data Intelligence and Security, ICDIS 2018. – P. 194-199.
[9] Ryu S., Kang Y.-J., Lee H. A study on detection of anomaly behavior in automation industry // International Conference on Advanced Communication Technology, ICACT 2018-February. – P. 377-380.
[10] Li Q., Liu P. Detecting user behavior anomalies in communication networks // 2nd IEEE International Conference on Cloud Computing and Big Data Analysis, ICCCBDA 2017. – P. 384-388.
[11] Ohori R., Torii S. Suspicious user detection based on file server usage features // Advances in Intelligent Systems and Computing, 612. – P. 467-470.
[12] Goldberg H.G., Young W.T., Memory A., Senator T.E. Explaining and aggregating anomalies to detect insider threats // Proceedings of the Annual Hawaii International Conference on System Sciences, 2016. – P. 2739-2748.
[13] Nazarov A.O., Anikin I.V. Clusterization of objects with fuzzy parameter's values // Proceedings of 11th IEEE Conference Dynamics of Systems, Mechanisms and Machines, Dynamics 2017.
[14] Anikin I.V., Gazimov R.M. Privacy preserving DBSCAN clustering algorithm for vertically partitioned data in distributed systems // Proceedings of 2017 International Siberian Conference on Control and Communications, SIBCON 2017.
[15] Garaeva A., Makhmutova F., Anikin I., Sattler K.-U. A framework for co-location patterns mining in big spatial data // Proceedings of 2017 20th IEEE International Conference on Soft Computing and Measurements, SCM 2017.
[16] Katasev A.S., Kataseva D.V., Emaletdinova L.Yu. Neuro-fuzzy model of complex objects approximation with discrete output // Proceedings of 2nd International Conference on Industrial Engineering, Applications and Manufacturing, ICIEAM 2016.
[17] Salakhutdinov R.Z., Ismagilov I.I., Rubtsov A.V. A neural-fuzzy approach to economic data classification // International Conference on Fuzzy Sets and Soft Computing in Economics and Finance. Proceedings. Volume II, 2004. – P.394-400.
[18] Liu P., Yao Z., Yin J. Improved decision tree of C 4.5 // Journal of Tsinghua University, 46. – P. 996-1001.
[19] Katarya R., Gangwar V., Jaisia I. A Study on different data mining classifiers // Proceedings of 2018 International Conference on Computer Communication and Informatics, ICCCI.

 © 2018 The Author (s); Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

[20] Thilina A., Attanayake S., Samarakoon S., Edirisinghe T., Krishnadeva K. Intruder detection using deep learning and association rule mining // Proceedings of 16th IEEE International Conference on Computer and Information Technology, CIT 2016. – P. 615-620.
[21] Gao S., Alhajj R., Rokne J., Guan J. Mining sequential patterns with extensible knowledge representation // Intelligent Data Analysis, 15(6). – P. 889-911.
[22] Muthuselvan S., Soma Sundaram K. A survey of sequence patterns in data mining techniques // International Journal of Applied Engineering Research, 10(1). – P. 1807-1815.

 © 2018 The Author (s); Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495