user profile, intruder detection
computer security
user verification

Tomasz WESOLOWSKI[1], Przemyslaw KUDLACIK[1]

DATA CLUSTERING FOR THE BLOCK PROFILE METHOD OF INTRUDER DETECTION

The paper concerns the problem of computer security and intrusion detection. In particular the problem of detecting the masqueraders – intruders who pretend to be authorized users of computer systems. The article presents an improvement of the block profile method of intruder detection for computer systems. As a base for the work the block profiles were taken and the clustering of the block profiles data was performed. The aim was to detect the masqueraders within the test data. The problem of computer security is particularly important in medical applications. Activities of intruder can distort the data stored in medical databases and analyzed by the medical expert systems or decision support systems and thus make it difficult to support the diagnosis of diseases. For this reason it is necessary protect the medical computer systems from unauthorized access. This article presents the preliminary research and conclusions.

1. INTRODUCTION

There is almost no aspect of human life where the computer systems are not used. One of the most important role of computer systems is being a part of a medical health care and rescue system. They provide services used in management of health care facilities, they allow fast and accurate analysis of medical data and support the diagnosis and treatment of diseases [4].

As the computers became a tool of daily use the danger of intrusion got very high. It is very important to secure the computer systems from those intrusions. The risk is particularly high at the medical facilities due to the special nature of their activities. First of all, medical centers process both the personal data of patients and their medical records, which by law are of a sensitive nature and should be particularly protected. In addition, the risk increases because of the open nature of the areas where computers are located in medical facilities, causing an easier access of potential intruders. The reasons mentioned above indicate that the protection of medical institutions against cyber-attacks is quite a significant matter and the security expectations are high. In this case the security of computer systems and networks is a top priority task.

There is many different kinds of cyber-attacks [2], however masqueraders constitute the greatest threat [5]. A masquerader in computer systems is an intruder who pretends to be a legitimate user - usually it is an insider intruder. Such an intruder temporarily overtakes the role of an authorized user to impersonate this user for malicious purposes. Methods discussed in this paper were created to detect such an intrusions in off-line mode. This means that the users activity data is collected and only after some time the detection process is carried out - contrary to on-line mode intrusion detection systems (IDS) that analyze user activity on the fly, in real time.

[1] University of Silesia, Institute of Computer Science, Bedzinska 39, 41-200 Sosnowiec, Poland
email: tomasz.wesolowski, przemyslaw.kudlacik@us.edu.pl.

Subsequent sections of this paper describe the problem of masquerade detection and the data set used in the experiments, present the command and block profiles of computer system user, introduce the idea of block profile data clustering and finally present the conclusions obtained from the results of experiments.

1.1. STATEMENT OF THE PROBLEM

The problem of detecting masqueraders in computer systems is clearly defined by Schonlau et al. (SEA) [6]. The aim is to develop a method working in an off-line mode, on a closed set of data - but the intruders are outside of the data set. The task is only to detect masquerades in the provided data set. Identifying intruders is not a part of the task because the intruders are from outside of the tested data set. The main assumption is that the not authorized users (illegitimate users called also alien users) behave differently at the time of the attack than authorized (legitimate) users.

1.2. DATA SET DESCRIPTION

As a result of an experiment performed by SEA the data set containing information about the activities of computer systems users was delivered. System calls made by 70 users of UNIX operating system in different institutions [8] were collected. For each user that took part in the experiment 15,000 commands were recorded. Out of all tested users 50 were randomly selected as native (legitimate) users and the remaining 20 became intruders (masqueraders). The commands of each user were divided into 150 blocks of 100 commands each. Next, 150 blocks of commands for each user were divided into three groups of 50 blocks. The first group I (blocks 1-50) consists of original (not contaminated) data. The other two groups (II: blocks 51-100 and III: blocks 101-150) were contaminated randomly by alien blocks (intrusions) taken from the data sets of 20 intruders. Not contaminated blocks of the first group create the training data set. Detailed description of the experiment together with the data set itself and the information about alien blocks can be found in [6], [7].

2. THE COMMAND AND BLOCK PROFILE

The Command and Block Profiles were introduced in [1]. To allow a statistical analysis of the data set delivered by SEA it was necessary to change its form. This is why the sequence of each user's commands was reorganized into a data matrix. Both profiles are created for each user separately. First, the number of unique commands u was determined. Next, the incidence matrix X was designated, where the number of rows equals the number of command blocks for one user (constant equal to 150) and the number of columns equals u of this user. The element x_{ij} of X indicates the is the number of occurrences of the command no. j where $j = 1 \dots u$ in block no. i.

To create the user's first profile his matrix X is divided into three parts corresponding to the three groups of blocks I, II and III. From each part a row vector of size $1 \times u$, containing the frequency of each command in that part of the matrix, is calculated. The values of the resulting vectors are plotted on the graph where the x-coordinates are in the range $[1, u]$, these graphs constitute the Command Profile of the user. The examples of Command Profiles are presented in Fig. 1 (for a user not contaminated with alien blocks) and in Fig. 2 (for a user which data was contaminated by alien blocks).

It can be clearly noticed in Fig. 1, that the profiles of all three parts are similar so with high probability it can be stated the in the data of user no. 11 no alien blocks are present. On the contrary, in the profile of user no. 24 (Fig. 2), where part III of the profile clearly shows the increase in the use of one of the commands that was nearly not used in the training part I. The Command Profiles can be analyzed visually, but a formal way to establish similarities is required. In [1] the Block Profiles as a tool to compare Command Profiles are introduced. Block Profiles are computer separately for each part of data. To create a Block Profile first, for each data block no. i where $i = 1 \dots 50$ a vector v_i of length u

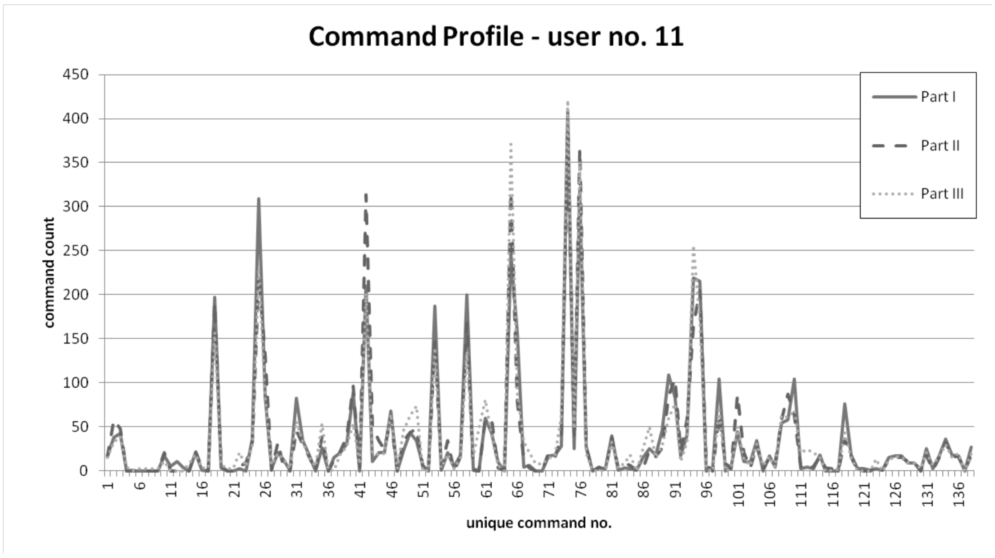

Fig. 1. Command Profile of uncontaminated user data.

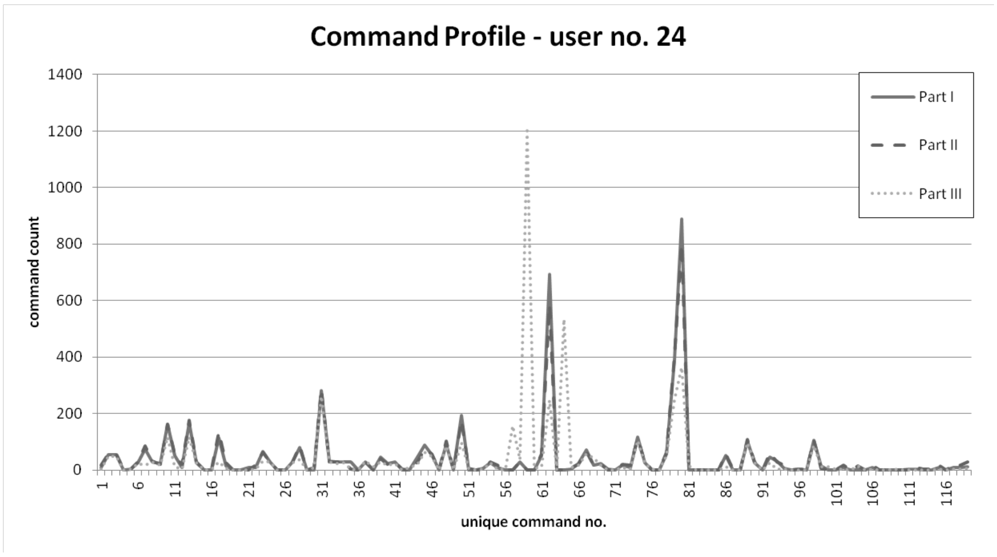

Fig. 2. Command Profile of a user data contaminated with alien blocks.

is calculated. These vectors contain the frequency of occurrences of each command in the i-th block. Next, for each block i, its squared Euclidean distance from the origin of the multivariate coordinate system is obtained by the following equation

$$d^2_{Eu}(i) = d^2_{Eu}(x_i - 0) = \Sigma^u_{j=1} x^2_{ij}, \tag{1}$$

where $i = 1, \ldots, 50$. At the end, a scatter plot constituting the Block Profile is created presenting the calculated distances d_{Eu} against the block number (as seen in Fig. 3 for uncontaminated user and in Fig. 4 for user data with alien blocks marked as discs).

The criterion for detection of alien block in the Block Profile method was the value of 5% of the largest distance values. This means that the blocks for which the distance belongs to the largest 5% of the highest values are considered alien blocks and constitute an intruder attack.

2.1. ANALYSIS OF THE PROFILES

The criterion used with Block Profile is too simple and it does not work for most of the cases due to the occurrence of alien blocks in the close relations to the normal blocks. In case of such a criterion it

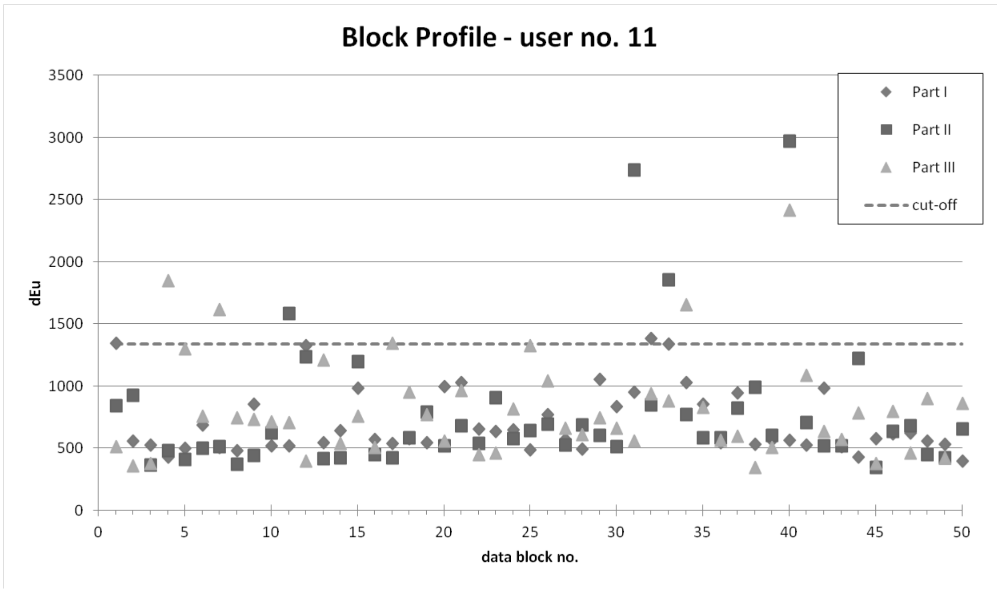

Fig. 3. An example of Block Profile for uncontaminated user data.

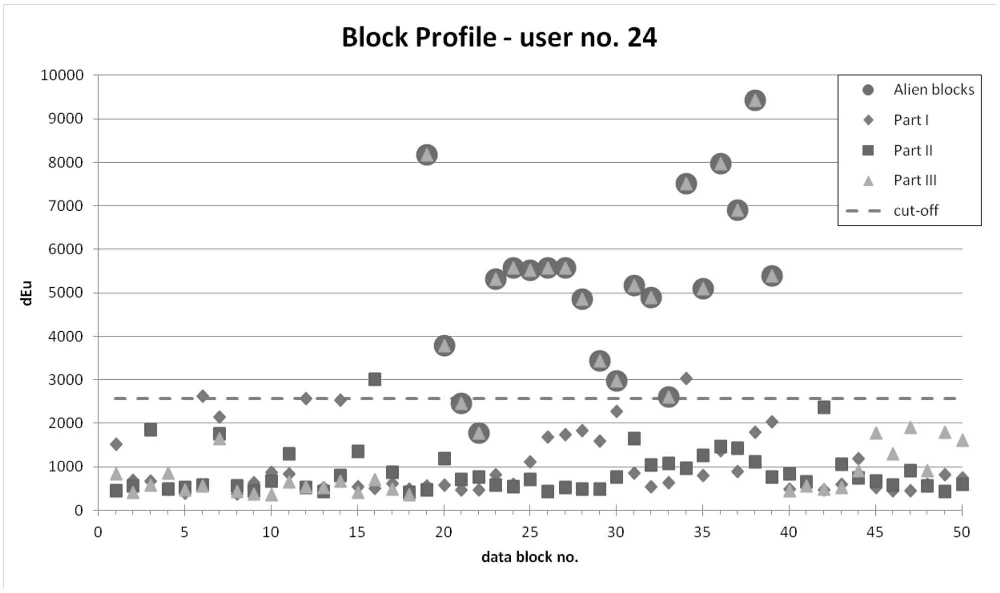

Fig. 4. Block Profile for user data contaminated with alien blocks.

is very difficult to determine the appropriate threshold for a given parameter since the number of false or undetected alarms depends on a level at which threshold was set. Another drawback of this method is that the plots for all three data parts are overlapping. In the data set the three parts are not related in such a way so the overlapping is artificial and it is not even necessary when using the introduced criterion. Also the x-axis of the plot is added artificially. When creating the profiles only the frequencies of the commands are used and the information about the commands and their order in a sequence is lost. Taking all above into consideration and due to the fact that in the Block Profile two dimensions are available instead of using the cut-off criterion the clustering to group the data can be used.

3. CLUSTERING DATA OF THE BLOCK PROFILE

The idea for analyzing the Block Profiles in order to find alien blocks by applying clustering is to look for the most separated group of elements. The most separated groups in a profile are considered as an intruder attack. In the described experiments a hierarchical clustering [3] was used. Hierarchical clustering methods need a specification of a dissimilarity measure between groups of observations. In the clustering process a hierarchical representations is created. The clusters at each level of the hierarchy are created by merging clusters at the next lower level. At the lowest level, each cluster contains a single value. At the highest level there is only one cluster containing all of the data. In this case the agglomerative approach (a bottom up type approach) was used in which at the beginning each value of the Block Profile starts in its own cluster and than in every subsequent step the pair of clusters with the smallest dissimilarity is merged hence at each level there one cluster less. In case of clustering the Block Profiles as a measure of dissimilarity the distance between the gravity centers of the clusters is used. This means that always the two closest clusters are merged.

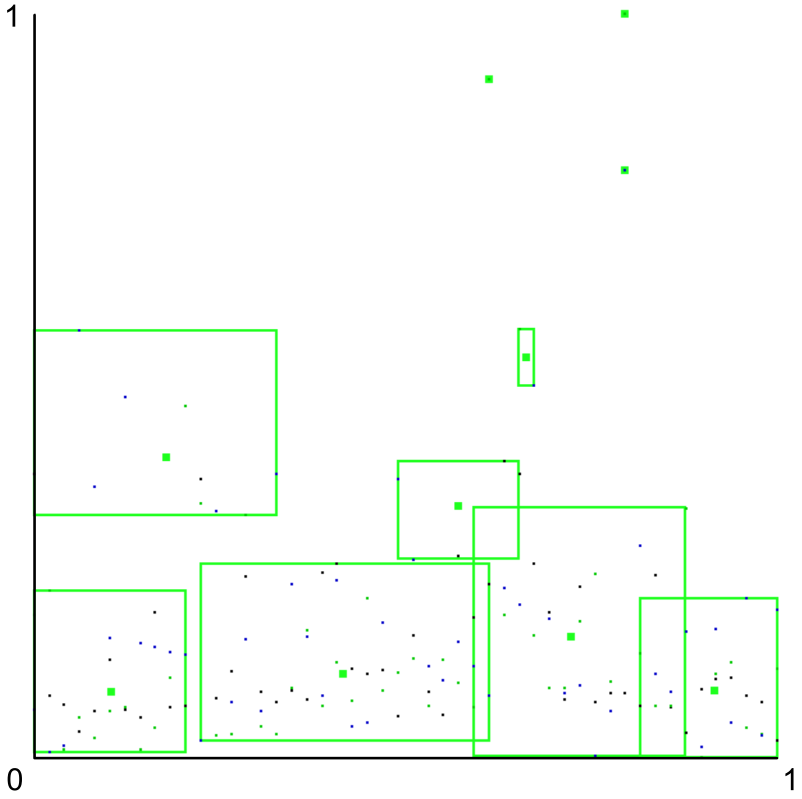

Fig. 5. Result of clustering for uncontaminated user no. 11 data.

In order to perform clustering the data of the Block Profiles was normalized to fit the range $[0, 1]$ in both dimensions. The algorithm stops either when the distance between the gravity centers of the clusters is greater than 0.3 or when there is a minimum of 10 clusters reached. These values were chosen empirically based on experimental observation. The result of the clustering can be observed in Fig. 5 and Fig. 6 for the uncontaminated and contaminated user respectively. The clustering approach seems to work better than original cut-off method. Unfortunately, the expected improvement level has not been achieved due to the issues with initial assumptions of the Block Profiles:

- the information about the command types is lost,
- the block profile contains overlapping local results (three subsequent, irrelevant groups).

4. MODIFICATIONS IN THE BLOCK PROFILE

To eliminate the drawbacks of Block Profiles some modifications were made. First, the range of the x-coordinates was changed to eliminate the artificial overlapping of the three data parts and to separate the training part I from the tested parts II and III. The analysis of the modified block profile (non-overlapping subsequent data) was performed and on the basis of observations it has been found that this solution may work better for. In Fig. 7 it can be observed that, in contrast to clustering performed on the original profile here alien blocks (middle sized rectangles) are grouped into separate clusters,

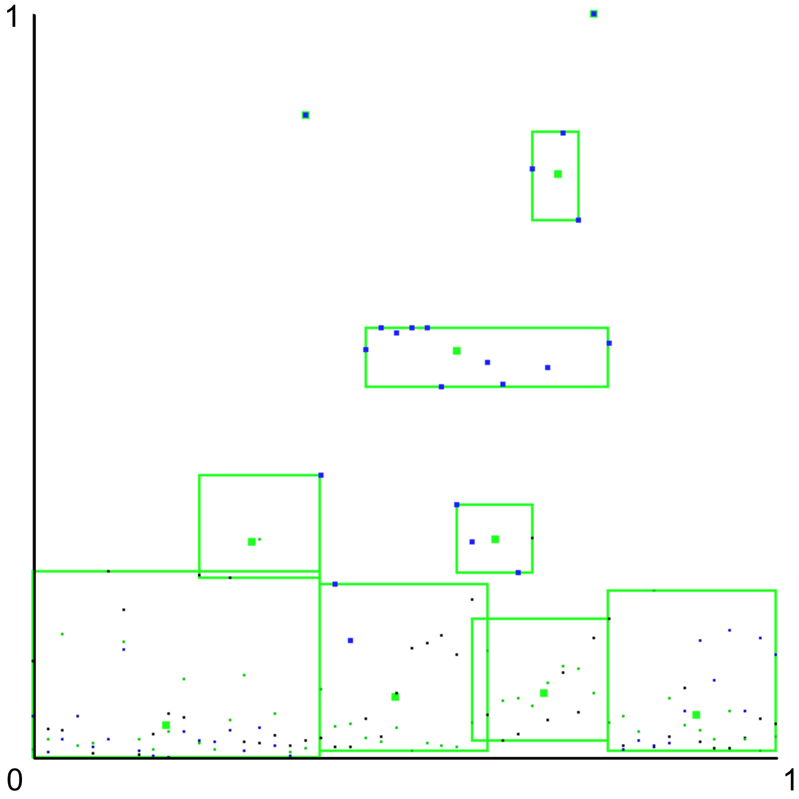

Fig. 6. Result of clustering for contaminated with alien blocks data of user no. 24.

and do not mix with normal blocks. Still the improvement level was not satisfying so another method was examined. The new method consists of two steps. First, the clustering is performed only on the training part of the data (part I). Second, clusters designated in the first step have been applied to the remaining two parts of the data. Using this method, the system first learns on the training set and then uses the learned rules to find the alien blocks by identifying the outlying values. Unfortunately for such a small amount of data the method does not work (as seen in Fig. 8).

5. CONCLUSIONS

The approach of Block Profiles works only with simple cases. For example when an intruder uses many times a single command that was not used by a legitimate user he is impersonating. When the case is more complicated problems occur. The original criterion is extremely simple and it does not work for most of the cases due to the occurrence of alien blocks in the close relations to the normal blocks. Statistical systems learn from the training data. If the training set is distorted, for example, by aggregating unauthorized activities carried out by an intruder, then these ineligible activities will be deemed as standard and a detection can become impossible. The use of statistical methods aims to detect abnormal situations, which suggest a potential danger. This means that the statistical systems are sensitive to any changes in user behavior. In real IDS it is a serious drawback, as in the case when the user temporarily takes over the duties of another user, and performs tasks that previously were not performed, the system will interpret this as an unusual and dangerous situation and report a false alarm.

In case of the Block Profiles method the second dimension is added artificially and the tree data parts are overlapping. The result is that possible alien blocks are mixed with the normal data blocks. Clustering improves the results but this method still outstands the expected detection level because it is based on the Block Profile and the Block Profile itself has to many drawbacks. Minor modifications made to the Block Profile and the method based on clustering also did not give the expected improvement.

The approach does not work because of information loss. Based on the observations made during experiments it is possible to propose some corrections while considering to build another profile:

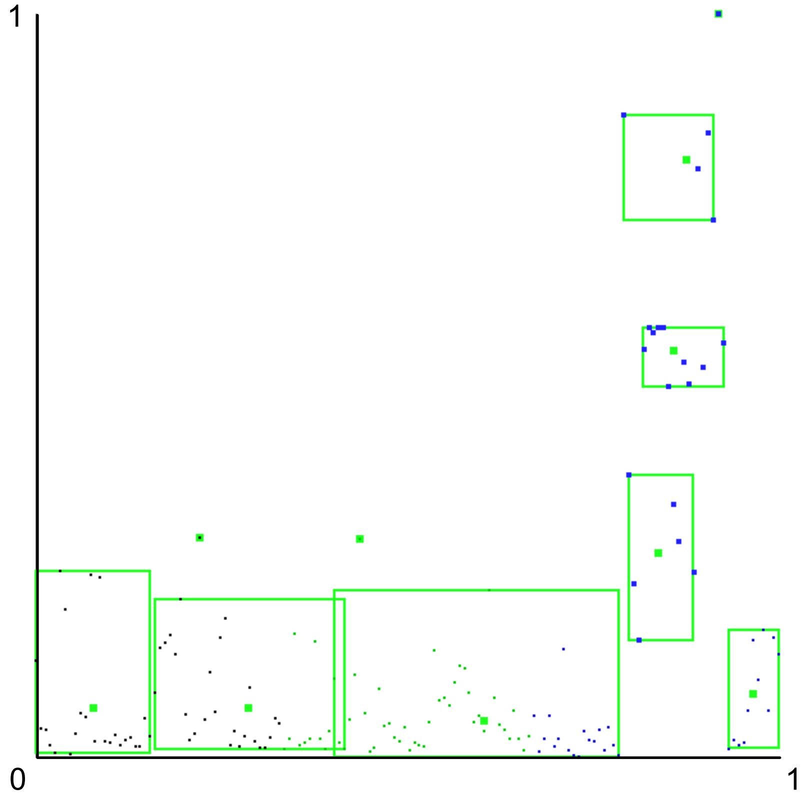

Fig. 7. Clustering modified profile of contaminated data of user no. 24.

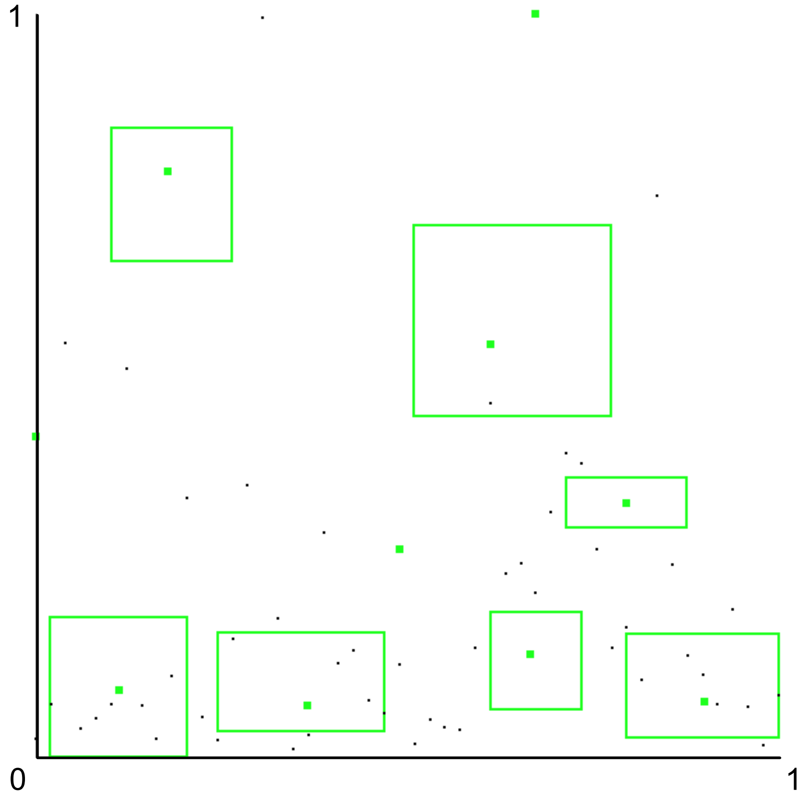

Fig. 8. Two step method applied to the profile of contaminated data of user no. 24.

- storing command names and their occurrence,
- changing the granularity of data set (100 commands in the block is quite a lot).

As a final conclusion we can say that due to the issues and disadvantages of discussed profiles described in the article the presented profile-based methods are too weak and it is unlikely to apply these methods into real IDS. Based on presented conclusions the work on the development of other, more efficient, method based on fuzzy inference was done and is described in a separate article.

BIBLIOGRAPHY

[1] BARTKOWIAK A. M., Block and command profiles for legitimate users of a computer network, Computer Information Systems - Analysis and Technologies - 10th International Conference, CISIM 2011, Proceedings. Communications in Computer and Information Science 245, Springer, 2011, pp. 295-304.

[2] DENNING D. E., Cyberspace attacks and countermeasures, In Internet Besieged D. E. Denning, P. J. Denning (eds), ACM Press, New York, 1997, pp. 29-55.

[3] HASTIE T., TIBSHIRANI R., FRIEDMAN J., The Elements of Statistical Learning (2nd edition), Springer-Verlag, 2009.

[4] PORWIK P., SOSNOWSKI M., WESOŁOWSKI T., WRÓBEL K., Computational Assessment of a Blood Vessels Compliance: A Procedure Based on Computed Tomography Coronary Angiography, LNAI, Springer, 2011, Vol. 6678/1, pp. 428-435.

[5] SALEM M. B., HERSHKOP S., STOLFO S. J., A survey of insider attack detection research, Insider Attack and Cyber Security: Beyond the Hacker, Springer, 2008, pp. 69-90.

[6] SCHONLAU M. et al., Computer intrusion: detecting masquerades, Statistical Science, 2001, Vol. 16, pp. 58-74.

[7] SCHONLAU M., Masquerading user data, http://www.schonlau.net.

[8] SCHONLAU M., THEUS M., Detecting Masquerades in Intrusion Detection Based on Unpopular Commands, Information Processing Letters 76, 2000, pp. 33-38.