Höglund A. J. and Hätönen K., 1998, Computer Network User Behaviour Visualisation Using Self Organising Maps, Proceedings of the 8th International Conference on Artificial Neural Networks (ICANN 1998), vol. 2, pp. 899-904.

Computer Network User Behaviour Visualisation Using Self Organising Maps

Albert J. Höglund
Nokia Research Center, Nokia Group
Helsinki, Finland
albert.hoglund@research.nokia.com

Kimmo Hätönen
Nokia Research Center, Nokia Group
Helsinki, Finland
kimmo.hatonen@research.nokia.com

Abstract

Computer systems are vulnerable to abuse by insiders and to penetration by outsiders. The amount of monitoring data generated in computer networks is enormous. Tools are needed to ease the work of system operators. Anomaly detection attempts to recognise abnormal behaviour to detect intrusions.

A prototype Anomaly Detection System has been constructed. The system provides means for automatic anomaly detection and user behaviour visualisation. The system consists of a data gathering component, a user behaviour visualisation component, an automatic anomaly detection component and a user interface. This paper is focused on the user behaviour visualisation component. This component uses large Self Organising Maps as a basis. The construction and the usage of the component is presented. Some discussion on comments from the test usage of the Anomaly Detection System is also provided.

1 Introduction

Computers and computer networks are becoming more and more important. The computer networks are normally protected from unauthorised usage by security mechanisms, such as passwords and access controls. However, if an abuser or intruder manages to bypass these security mechanisms and gains access to vital information, the potential loss is enormous.

The loss can be decreased by detecting intruders or abusers at an early stage. The two main intrusion detection techniques are rule-based misuse detection and anomaly detection. Rule-based misuse detection attempts to recognise specific behaviours that are known to be improper. That is, if a user follows certain intrusive patterns, he is classified as an intruder. Anomaly detection, on the other hand, attempts to recognise anomalous or abnormal user behaviour to detect intrusions. Anomalous or abnormal behaviour is suspected if the current behaviour deviates sufficiently from the

previous behaviour, which is assumed normal. Anomaly detection deals with behaviour that is not known in advance, while rule-based misuse detection deals with predefined violations. In this difficulty also lies the advantage of anomaly detection: it can namely be used to detect types of intrusions that have never occurred before. References on anomaly detection and misuse intrusion detection research can be found in [1, 2, 3, 4].

A prototype Anomaly Detection System for the UNIX environment has been constructed [5]. The system provides means for automatic anomaly detection and user behaviour visualisation. The system consists of a data gathering component, a user behaviour visualisation component, an automatic anomaly detection component and a user interface. This paper presents the user behaviour visualisation component.

The objective with the user behaviour visualisation was to visualise the user behaviour during a certain period in a simple way. Such a component is useful, since the amount of data generated by users in a computer network is enormous.

The amount of information is reduced by selecting a set of features that characterises the behaviour of the users in the network. This set of features should form a daily fingerprint of the network user, which means that it has to be selected carefully. Although the amount of data is reduced in the feature selection process, it is still difficult to compare and analyse the user behaviour. The visualisation problem is tackled using an approach, in which the Self Organising Map is used to visualise the user behaviour in two dimensions. The Self Organising Map is a neural network based on unsupervised learning and it is suitable for visualisation and interpretation of large high-dimensional sets of data.

2 Methods and Implementation

2.1 Software and Environment

The prototype Anomaly Detection System was built in the UNIX environment. The data gathering and the data processing are performed on separate servers. This enhances security and makes it more difficult to disturb the operation of the system.

The routines that build the prototype Anomaly Detection System have been coded using C and Perl and the Self Organising Maps are made using the procedures of SOM_PAK [6]. The user interface uses Netscape[1] to view the html pages generated by the system.

2.2 Data Gathering and Scaling

For a period of 400 days the user account logs of more than 600 users have been stored. The user account logs give information on the processes performed by the users. This information includes CPU-times, characters transmitted and blocks read. The selection of features describing the user behaviour is discussed in Section 2.3.

[1] The Netscape browser can be downloaded from http://www.netscape.com/

Since the magnitude of the features varies greatly, logarithmic or linear scaling was considered necessary. The features were scaled according to (1) and (2), where *f* is the feature in question. The division by the maximum scales the parameters to the range [0, 1]. Histograms were studied in order to determine which one of the two scalings was more suitable for the feature in question.

$$f_{i\_Log-scaled} = \frac{\ln(f_i + 1)}{\max_i[\ln(f_i + 1)]}, \quad f_{i\_Lin-scaled} = \frac{f_i}{\max_i[f_i]} \qquad (1), (2)$$

2.3 Feature Selection

Feature selection provides a means of reducing the enormous amount of data generated by computer network users. The feature selection problem can be stated as follows: Objects are described with a large set of features. The objective is to find a subset of features that distinguishes the objects from each other as well as possible.

Features characterising the user behaviour during a period of 24 hours are used in the user behaviour visualisation component. Firstly an initial feature set with 34 features was derived. Features describing CPU-time and transmitted characters for different services were included, but also session, process and login information. The initial feature set was reduced to a set of 16 features. This was achieved by omitting features with strong linear dependency to other features and by omitting very noisy features. The linear correlation was checked using a correlation test and the noisy features were found by closely examining the variances. Careful consideration was used in the feature selection process.

2.4 User Behaviour Visualisation

The idea is to use Self Organising Maps (SOM) to visualise the user behaviour during a certain period in a simple way. The SOM is an effective tool for visualisation of high-dimensional data [7, 8]. The principal goal of the SOM is to transform an incoming signal pattern of arbitrary dimension into a one- or two-dimensional discrete map, and to perform this transformation in a topologically ordered fashion. The algorithm and detailed theory on the SOM can be found [7].

The user behaviour visualisation component uses two-dimensional SOM:s. The maps are constructed using the whole data set for the whole period, which means the data for all the users for all the days in the period. A large number of tests indicated that maps of size 18x14 give sufficient accuracy. The type of lattice used is hexagonal (six neighbours) and the neighbourhood function type used is bubble [6, 7]. In this paper the map has been labelled with the user number and the number of "hits" on the neuron. These labels are separated by a "_". The maps are visualised using the U-matrix method [9, 10, 11]. In this method the neurons are marked with dots and the distances between them are described with greyscales. The darker the cell between two neurons, the greater the distance between them. In addition, the user behaviour visualisation component provides real values for the neurons of the map, a connection to the real data and feature statistics.

3 Using the User Behaviour Visualisation Component

The user behaviour visualisation component of the Anomaly Detection System prototype has been in test usage during a period of several months. The following user behaviour clustering examples illustrate the use of the user behaviour visualisation component.

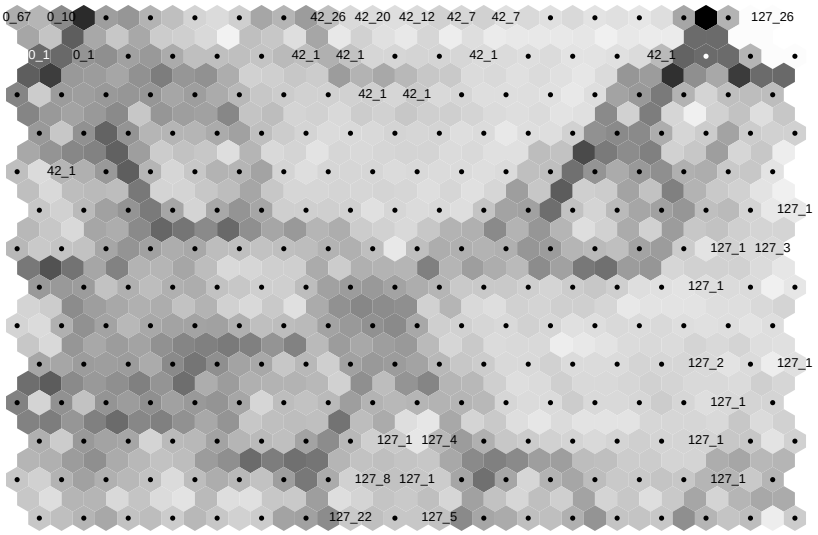

Figure 1 Map of size 18x14 trained with the whole set of data and labelled with the usage of user 0, 42 and 127

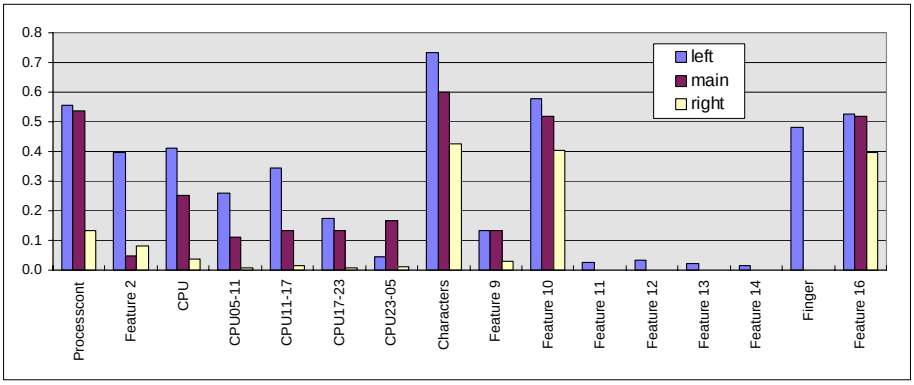

Figure 2 Feature distribution for user 42 main usage cluster compared with the deviation on the right and the deviation on the left.

The behaviour of three users during a period of 79 days is visualised in Figure 1. The behaviour of user 42 is quite nicely clustered. Figure 1 shows that user 42 has

one main usage cluster in the upper middle of the map, but one notices that there are two deviations from this cluster, one on the right and one on the left. An explanation to these deviations is given in Figure 2. In this figure the feature distribution of the neuron with 20 hits from the main usage cluster is compared with the feature distribution of the deviations on the left and on the right. The network usage of the deviation on the right is lighter than normal with fewer processes, less CPU-time and fewer characters transmitted. This deviation can be explained with a breakdown in the network. The deviation on the left, on the other hand, is a bit anomalous. The usage is heavier than normal, especially the CPU-times in the morning and in the afternoon are bigger than normal. Another anomalous thing with the deviation on the left is the high usage of the finger service compared with no use at all for the normal behaviour.

The behaviour of hundreds of users have been analysed during the test usage. There is for example a well bounded no usage cluster in the upper right corner of the map in Figure 1. Users that rarely use the computer network are mostly mapped to this cluster. Figure 1 shows that user 127 has 26 days with no usage during the period of 79 days. Another well-bounded cluster is located in the upper left corner of the map. A system demon id with user number 0 is mapped to this cluster (see Figure 1). The behaviour of this id can be considered very regular since its behaviour is mapped to only four neurons on the map. A misuser or intruder using the id of the system demon would be very critical, since it has greater privileges in the network than the normal users. Deviations in the behaviour of the system demon id can easily be noted using the user behaviour visualisation component.

There are also users that are mapped to two clearly separated usage clusters, which means that the users have two working modes. An example of this is user 127. Figure 1 shows that the behaviour of user 127 is clustered to three clearly separated clusters. One is the no usage cluster and the two others are the normal working modes. These modes can be analysed further using the same procedures as with user 42 above. The phenomenon with two working modes may, for example, originate from the fact that the users work in more than one project. There are of course some users whose behaviour is very irregular and their behaviour is therefore not so nicely clustered on the map.

4 Conclusions and Discussion

The initial feedback from the test usage of the Anomaly Detection System has been quite positive. Comments like “The user behaviour visualisation component gives a quick overview of the user behaviour” were quite encouraging. The test usage feedback also indicated that the component is practical when analysing user behaviour that has been reported anomalous. Section 3 showed how the Self Organising Map in the user behaviour visualisation component can be used to analyse user behaviour. The user feedback also included comments on necessary improvements. Better connections to the real data for further analysis were suggested and have now been implemented. Improvements in the labelling of the map and map region classification have also been suggested and will be implemented in the future.

The general impression of the authors is that the Self Organising Map provides a good method for reducing the dimensions of the data and for comparing and visualising the behaviour of network users. Examples and test usage give just indications of the performance of the user behaviour visualisation component, though. Simulation experiments for further evaluation of both the user behaviour visualisation component and the automatic anomaly detection component will therefore be performed. A publication on the automatic anomaly detection component is also under preparation.

References

[1] Javitz H S, Valdes A, Lunt T F, Tamaru A, Tyson M, Lowrance J. Next generation intrusion-detection expert system (NIDES): Statistical algorithms rationale and rationale for proposed resolver. Technical report, Computer Science Laboratory, SRI International, Menlo Park, California, The USA 1993.

[2] Kumar S, Spafford EH. A pattern matching model for misuse intrusion detection. In Proceedings of the 17th National Computer Security Conference, October 1994, pp 11-21.

[3] Lankewicz L B, A nonparametric pattern recognition approach to anomaly detection. Doctoral Thesis, Tulane University, 1992.

[4] Lunt T. F. A survey of intrusion detection techniques, Computers and Security 1993; 12(4):405-418

[5] Höglund A. An Anomaly Detection System for Computer Networks. Master's thesis, Helsinki University of Technology, Helsinki, 1997.

[6] Kohonen T, Hynninen J, Kangas J, Laaksonen J. Manual of SOM_PAK, The Self-Organising Map Program Package, Version 3.1, April 7, 1995, http://nucleus.hu-t.fi/nnrc.html.

[7] Kohonen T. Self Organising Maps. Second Edition. Springer-Verlag, Heidelberg 1997.

[8] Neural Networks Research Centre & Laboratory of Computer and Information Science. Triennal Report 1994-1996. Helsinki University of Technology 1997.

[9] Iivarinen J, Kohonen T, Kangas J, Kaski S. Visualising the clusters on the Self Organising Map. Multiple Paradigms for Artificial Intelligence (SteP94), 122-126. Finnish Artificial Intelligence Society, 1994.

[10] Kraaijveld M A, Mao J, Jain A K. A non-linear projection method based on Kohonen's topology preserving maps. Proceedings of the 11th International Conference on Pattern Recognition (11ICPR), 41-45, Los Alamitos, CA. IEEE Comput. Soc. Press, 1992.

[11] Ultsch A, Self organised feature maps for monitoring and knowledge acquisition of a chemical process. Gielen S, Kappen B, editors, Proceedings of the International Conference on Artificial Neural Networks (ICANN93), London. Springer-Verlag, 1993, pp 864-867.