A Correlation Framework for Continuous User Authentication Using Data Mining

H. Singh[1], S.M. Furnell[2], P.S. Dowland[2], B. Lines[2] and S. Kaur[3]

[1]School of Computing Science, Middlesex University, London, UK
[2]Network Research Group, School of Computing, Communications and Electronics, University of Plymouth, Plymouth, UK
[3]Institute of Biological Sciences, University of Malaya, Kuala Lumpur, Malaysia
email: info@network-research-group.org

Abstract

The ever-increasing security breaches by both external and internal intruders highlight the lack of security measures in many current systems. Extensive work has been carried out to address this problem, for example by enhancing the initial login stage in order to overcome the security flaws of traditional authentication methods. However, in the event that an unauthorised user compromises a systems initial authentication, the user is in the position to do virtually anything without being further challenged. This has caused interest in the concept of continuous authentication during a user's active session based upon their behaviour characteristics, which inevitably involves the analysis of vast amounts of data. Whereas most reported work in this area uses statistical approaches to model the temporal regularities exhibited by users, this paper presents a series of comparative studies carried out using Data Mining techniques and algorithms. It presents the result of the analysis carried out and discusses a proposed systematic correlation framework for continuous user authentication using the Data Mining methodology adopted in the comparative studies. This paper shows how the correlation framework could be used to automate the analysis of the generated audit data as well as the processes involved in authenticating users in a networked environment.

Keywords

Authentication, Biometric, Data Mining, Behavioural Profiling, Intelligent Data Analysis, Keystroke Analysis.

1. Introduction

The increasing security breaches revealed in recent surveys (CSI 2003; NCC, 2000) and security threats reported in the media (Betts, 2000; Ward, 2000) reaffirms the lack of current security measures in IT systems. While most reported work in this area (Miller, 1994) has focussed on enhancing the initial login stage in order to counteract against unauthorised access, there is still a problem detecting when an intruder has compromised the front line controls. This could pose a serious threat since any subsequent indicator of an intrusion in progress could be quite subtle and may remain hidden to the casual observer. Having passed the frontline controls and having the appropriate access privileges, the intruder may be in the position to do virtually anything without further challenge. This has caused interest in the concept of continuous authentication, which inevitably involves the analysis of vast amounts of data. Although there has been some research (Seleznyov *et al.*, 2002) in applying the concept of continuous authentication, most of the reported work in this area uses statistical based approaches to model the temporal regularities exhibited by users for the authentication process.

This paper presents some comparative studies carried out using Data Mining (DM) techniques and algorithms. In section 2, we provide details of the comparative studies and results of the analysis carried out using DM algorithms on the generated behavioural data. Although some reported work has been carried out to analyse traffic data using DM (Warrender *et al.*, 2002; Lee & Stolfo, 2000), none to the knowledge of the authors has investigated the feasibility and effectiveness of learning techniques (e.g. neural network, machine learning, etc.) for the analysis of audit trails. The potential of using DM for the purpose of continuous authentication is further substantiated with a comparative study using a statistical approach for detecting deviation from a user's historical keystroke profile captured under a multi-tasking windowed environment. Since the initial work by Jobusch and Oldehoeft (1989), the use of keystroke analysis has been further investigated using Bayes classifiers (Bleha *et al.*, 1990) and statistical approaches (Joyce & Gupta, 1990) to analyse the keystroke data. It is, therefore, considered that there exists the scope for using DM techniques. The paper then proceeds to describe a proposed correlation framework for continuous user authentication using the DM methodology adopted in the comparative studies. This is considered novel since no work to the author's knowledge have proposed a correlation framework for continuous user authentication using DM.

2. Comparative Study on Behavioural Profiling Using Data Mining

We use DM to extract latent patterns or models of user behaviour from the collected audit trail. This is then reflected in the DM algorithm classifiers (e.g. through rule induction) to recognise deviation, if it occurs, from normal use. This approach is based on the assumption that a user's behaviour has regularity and that using the classifiers this behaviour can be modelled. Using this analogy, anomalous behaviours can then be categorised as a possible unauthorised user or use of that system. The audit trail data analysed was collected from networked computers on a participating local area network (LAN) using an independent agent installed locally in order to audit user interaction with the system. This is based on the assumption that users performing their regular tasks will impose similarly regular demands upon system resources. A number of system parameters were monitored and logged including resource usage and process-related information such as creation, activation and termination. Similar system features have been used in other published work (Lunt, 1990), however, this focused on statistical and neural network analysis. A user's behaviour profile can be uniquely identified by: <user name, absolute time, date, hostname, event1,…, event*n* >, which is the semantic used for the audit trail where, eventsn denotes the system features being monitored.

The methodology used is derived from the four main activities of DM; selection, pre-processing, data mining and interpretation (Figure 1). The collected audit trail is split into various sample sizes. These subsets form the target data sets, which will undergo the analysis to identify patterns and to test specific hypotheses. The cleaned data, containing both categorical and numerical data, is then subjected to analysis by the DM algorithms. There are a wide variety of DM techniques available, each of which performs more accurately over certain characteristic data sets (e.g. numerical or categorical) and is also relative to the number of variables or attributes and classes. The Intelligent Data Analysis (IDA) Data Mining Tool (Singh *et al.*, 1999) is used to analyse the sample data sets which incorporate algorithms from the fields of Statistical, Machine Learning and Neural Networks. Six algorithms, k-NN, COG, C4.5, CN2, OC1 and RBF were chosen for this investigation. For the purpose of this work, the data sets were split into ratios of 9:1, 8:2 and 7:3, hence into two

parts, which is a commonly used technique known as train and test. The algorithm or classifier is initially subjected to the training set and then the classification accuracy is tested using the unseen data set or testing set. The results give an indication of the error rate (or false positives) and the overall classification accuracy of the trained algorithms.

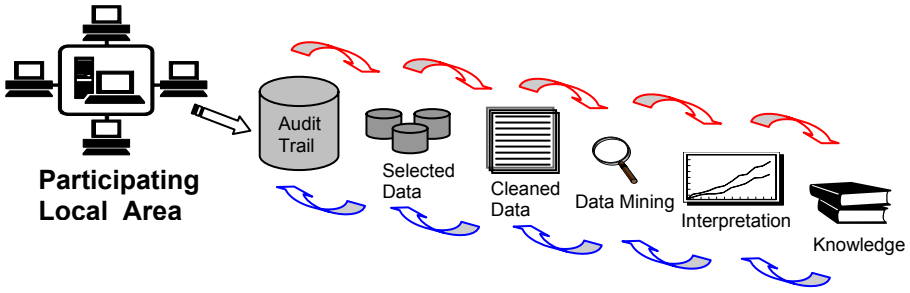

Figure 1. Methodology for Behavioural Profiling

2.1 Discussion of results

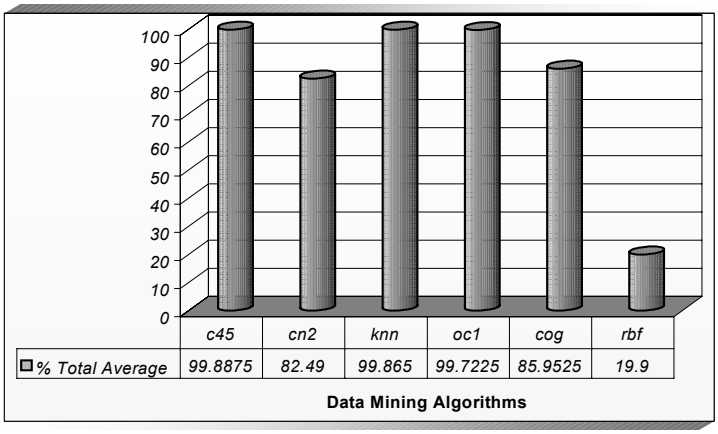

Figure 2. Total percentage average classification accuracy of selected DM algorithms

The initial results obtained from the analysis (Figure 2), suggest that Machine Learning and Statistical-based algorithms are better for these types of data sets. C4.5 and OC1 decision tree based algorithms in particular, out performed the CN2 rule-based and RBF algorithms. The classification accuracy obtained, using k-NN in comparison to C4.5, shows some significance for further investigation despite the slower classification times observed. Amongst the statistical algorithms, k-NN faired better then COG but is slower in comparison to the classification times observed. The classification accuracy obtained overall indicates that RBF classification accuracy is inversely proportional to the sample size. These results support other reported work (Michie *et al.*, 1994). In addition to the consistency in classifying the data sets and the overall average classification accuracy, our initial investigations also identified that C4.5 has overall quicker train and test time and outputs explicit rules.

3. Comparative Study on Keystroke Data Analysis Using Data Mining

Keystroke analysis is an example of a biometric that uses inter-keystroke latencies (time between keystrokes) to differentiate between users. While in the previous section the aim was to show the feasibility and effectiveness of DM learning algorithms in building temporal

regularities in user behaviour, this section is intended to build upon the findings by comparing against statistical approaches. In order to collect the required data, an independent agent installed locally on the networked computers was used for acquiring keystroke notifications across all applications running within a users' active session. A total of ten users were profiled out of which only 4 users (who provided the largest profiled data sets) were selected. The audit trail generated contained the following attributes: <first character, second character, digraph latency>.

Full details of the statistical method used are detailed in (Dowland *et al.*, 2001). Therefore only the relevant results are presented here. A summary of the profiles generated by the statistical method is shown in Table 1.

User	Unique Digraph Pairs	Filtered Digraph Pairs	Average Inter-keystroke Time
User A	466	122	151ms
User B	405	51	145ms
User C	412	89	206ms
User D	461	127	162ms

Table 1. Summary of user profile statistics

Once a user profile was generated, the profile was evaluated by comparison with the users' raw keystroke data. This allowed the test profile to be evaluated against the users' own data (to test the False Rejection Rate – FRR) and against other users' keystroke data (to test the False Acceptance Rate – FAR).

3.1 Discussion of results

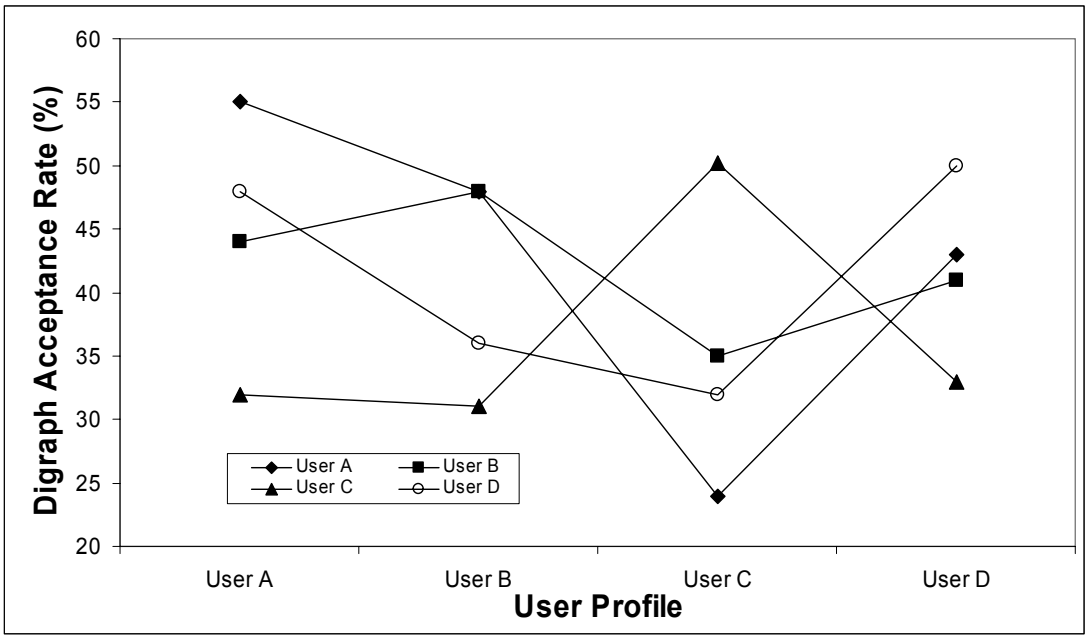

Figure 3. User profile comparisons

When viewing the preliminary results (Figure 3) if we consider the four users A, B, C and D and follow the vertical columns of data, we can see a clear peak for each user's data when

compared with their own profile. This is most noticeable for user C where a significant peak is observed (50% of all digraphs accepted) compared with 35% when user B's digraph data was tested against the same profile. Although there was a clear correlation between user C's profile and data, if we consider user A, there was a high FAR for data from users D and B (impostors) when compared with user A's profile. We can also see that in user B's profile the impostor "user A" achieved the same acceptance rate (48%).

3.2 Data Mining approach

The methodology used to analyse the raw keystroke data using DM followed a similar principle to that described in section 2. For the purpose of this work, the data sets were split into a ratio of 9:1. The algorithm or classifier is subjected initially to the training set and then the classification accuracy is tested using the unseen data set or testing set. The results give an indication of the error rate (or FAR) and the overall classification accuracy of the trained algorithms. The percentage acceptance rate obtained is encouraging (Figure 4), when considering the acceptance rate achieved for the highest algorithm is 53%. This is in consideration of the time factor involved in comparison to the statistical approach and the amount of domain expertise input to the process, which only resulted in an absolute difference of 2% from the highest percentage acceptance rate (i.e. 55%) obtained in the statistical analysis. Furthermore the acceptance rate obtained increases proportionally (unlike the statistical approach which is restricted in the sample size analysed), except for the COG and RBF algorithms. This is important when considering the size of data being analysed and hence eliminates the ad-hoc approaches adopted using traditional statistical methods. The initial results suggest that Machine Learning (OC1 and C4.5) and Statistical (k-NN) based algorithms are suitable for these types of data sets. Despite the results, more work needs to be carried out in order to correlate the results to a specific or group of algorithm(s), in order to obtain a higher percentage of classification accuracy. Nevertheless it is clear from the comparative study carried out that DM algorithms have the potential to automate the process of discovering the temporal regularities from the data sets, which would otherwise rely heavily upon intuition and experience in building this model using other approaches (e.g. statistical approach). Furthermore the methodology and algorithms provides the foundation, which could be integrated into a correlation framework as presented in the following section.

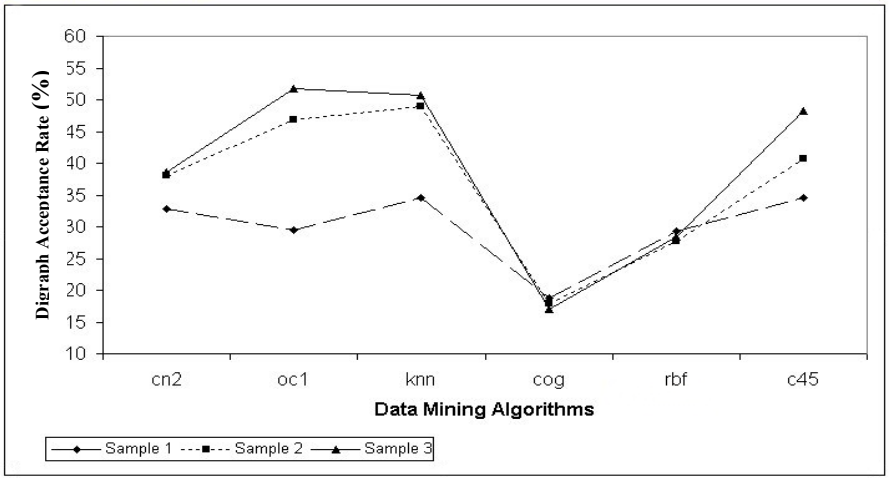

Figure 4. Varying sample sizes with fixed number of classes and attributes

4. Correlation Framework

The approaches investigated for user behaviour and process activity profiling could be used as the basis to provide user identification and authentication. Our initial results show the potential of developing and integrating the DM techniques investigated into a correlation framework, which could be integrated into an operating system user authentication scheme (Figure 5). The concepts behind the correlation framework are in some ways similar to the principles of inductive reasoning (Durkin, 1994) where the goal is to arrive at a decision (i.e. if deviation is occurring) from a limited set of information (i.e. behaviour indicative data) available due to the inherent problem of gleaning specific information from audit trails. The key aspects of this design are defined in the sections that follow.

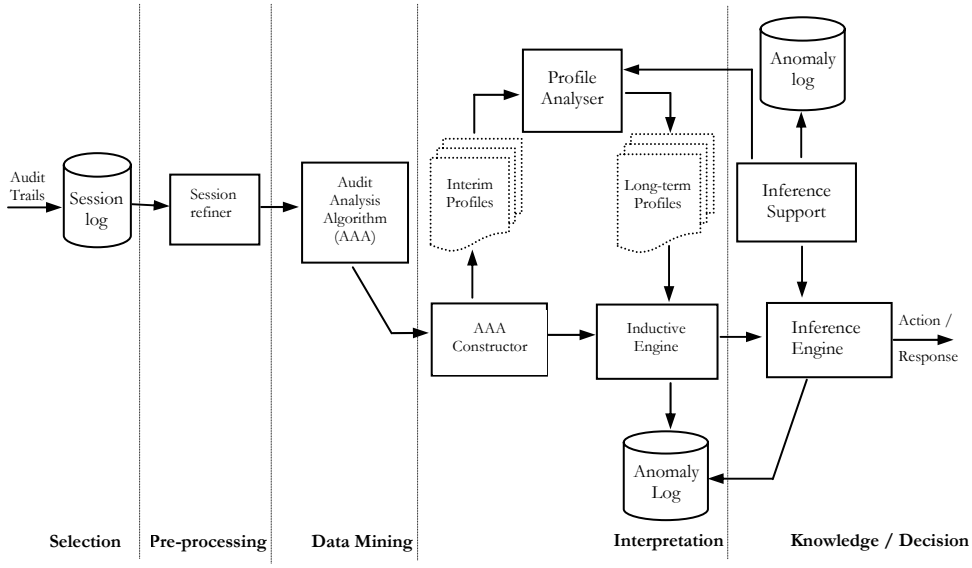

Figure 5. Correlation framework for continuous user authentication using Data Mining

4.1 Session Log

The Session Log would provide a temporary storage to the generated audit trails. The data stored would be restricted to only relevant data pertaining to the behavioural data (e.g. resource usage data, keystroke data, etc.), although it can potentially be a source for a permanent record to provide evidential support should the need transpire. This would enable a reduced access time for the Audit Processing module to select the relevant features for analysis in the later stage.

4.2 Session Refiner

The Session Refiner prepares the target data set prior to analysis. Typically this may involve converting the data into an acceptable format, annotating the beginning of each message (e.g. time stamp, absolute time, etc.) or may involve processes to quantify continuous features, e.g. to generate the audit trail semantic. These stages can be used to constrain the search space and make patterns or relationships in the audit trails more visible in the later stages of the correlation process.

4.3 Audit Analysis Algorithms

The Audit Analysis Algorithms (AAA) incorporate the DM algorithm(s) that will be used to identify system features, patterns and latent trends for classifying user behaviour. Data features such as correlation between adjacent or frequent sequential patterns of user behaviour will be analysed. The information gleaned from using these algorithms will be used to identify the temporal regularities of a user's behaviour, which will be reflected in the user's profiles in the latter stages.

4.4 AAA Constructor

The AAA Constructor would refine the inferred association rules or classification rules from the AAA engine. The various types of patterns exhibited in the data would be cleaned (i.e. removing redundant data), combined, and transformed into an understandable syntax. This will later be stored in the Interim Profile, which would provide a temporary repository.

4.5 Profiler Analyser

The concepts behind the Profiler Analyser are adapted from the IMS Profile Refiner (Furnell, 1995) and would have a similar functionality in the proposed correlation framework. Similarly the audit trails generated would be optimised as input to the Inductive Engine to detect deviation from normal use and as a source for updating user profiles, which will inevitably change over time. Depending on how often similar patterns are exhibited by user(s), these changes will be reflected in the Long-term Profiles repository.

4.6 Inductive Engine

The Inductive Engine would enable the detection of any deviation occurring on the system. The Long-term Profiles of users would be compared against the generated audit trails to detect for anomalies. Anomalies detected would be stored in the Anomaly Log, which would provide the basis for detected deviations to be further analysed in order to reduce the probability of false positives prior to reporting the conclusion inferred through the Inductive Engine.

4.7 Inference Engine

The Inference Engine would be used to identify recurring valid behavioural patterns that are being flagged as anomalies. These would be filtered out in order to reduce the potential of high false positive errors by correlating previously known anomalies logged in the Anomaly Log to the current active anomaly detected and from known information input through the Inference Support component.

4.8 Inference Support

The Inference Support would be used to improve the inferred facts from other sources. System Administrators or Security Officers could input this information (e.g. public holidays, staff on sick leave, etc), which would otherwise take a longer time, through normal circumstances, to infer and thus detect anomalies occurring. Furthermore it would be used to provide input to the Profile Analyser where deemed necessary for instance, to disable profiling when a user is away as a countermeasure against the possibility of an unauthorised

user introducing new temporal behaviour which would effect the legitimate user's profile. It would also enable any modifications (i.e. removing or adding anomalies) or maintenance required in the log files of the Anomaly Log.

5. Discussion and Conclusion

While there is a tendency to equate complex statistical analysis with correlation or the detection mechanism, this paper has presented the results to date from the comparative studies carried out using DM. The methodology used and the classification accuracy obtained in this initial investigative work suggests that DM techniques could be integrated into a correlation framework for continuous authentication. The high classification accuracy obtained and fast response time exhibited in classifying the user behaviour by some of the DM algorithms, when considering the vast amount of audit trail analysed, further demonstrates the potential of applying DM techniques within a real-time application. Whereas previous work in this area has been focussed on developing the DM algorithms for domain specific problems, no work to date has integrated these techniques into a correlation framework. The methodology developed in analysing the generated audit trails, which is advocated by the proposed correlation framework, has the potential to provide an important contribution to the development of a correlation framework for the purpose of continuous user authentication.

6. References

Betts B. (2000), "Digital Forensic: crime scene", *Information Security Magazine*, March, http://www.infosecuritymag.com/articles/march00/cover.shtml

Bleha S., Silvinsky C. and Hussein B. (1990), "Computer-Access Security Systems Using Keystroke Dynamics", *Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 12

Computer Security Institue (2003), "*2003 CSI/FBI Computer Crime and Security Survey*", http://www.gocsi.com/

Dowland P.S., Singh H. and Furnell S.M. (2001), "A Preliminary Investigation of User Authentication Using Continuous Keystroke Analysis", *Proceedings of the IFIP 8th Annual Working Conference on Information Security Management & Small Systems Security*, Las Vegas, 27-28 September

Durkin J. (1994), "*Expert Systems Design and Development*", Prentice Hall, ISBN 0-02-330970-9, pp. 90-130

Furnell S.M. (1995), "*Data security in European Healthcare Information Systems*" PhD Thesis, University of Plymouth

Jobusch D.L. and Oldehoeft A.E. (1989), "A Survey of Password Mechanisms: weaknesses and potential improvements. Part 1", *Computers & Security*, pp. 587-603

Joyce R. and Gupta G. (1990), "Identity Authentication Based on Keystroke Latencies", *Communications of ACM*, vol. 33, February

Lee W. and Stolfo S. (2000), "A Framework for for Intrusion Detection Systems", *ACM Transactions on Information and System Security*, vol. 3, no. 4, November

Lunt T.F. (1990), "IDES: an intelligent system for detecting intruders", *Proceedings of the Computer Security, Threat and Countermeasures Symposium*, Rome, Italy, November

Michie D., Spiegelhalter D.J. and Taylor C.C. (1994), "*Machine Learning, Neural and Statistical Classification*", Ellis Horwood, ISBN 0-13-106360-X, pp. 136-141

Miller B. (1994), "Vital Signs of Identity", *IEEE Spectrum*, February

National Computing Centre (2000), "*The Business Information Security Survey (BISS 2002)*", http://www.ncc.co.uk/ncc/biss2000.pdf

Seleznyov A., Mazhelis O. and Puuronen S. (2002), "Anomaly Intrusion Detection System Based on Online User Recognition", *Proceedings of the Third International Network Conference (INC 2002)*, Plymouth, UK, 16 – 18 July

Singh H., Burn-Thornton K.E. and Bull P.D. (1999), "Classification of Network State Using Data Mining", *Proceedings of the 4th IEEE MICC & ISCE '99*, Malacca, Malaysia, vol. 1, pp. 183-187

Ward M. (2001), "Web Warning Centre in Net Attack", *BBC News Web Site*, 24th May 2001, http://news.bbc.co.uk/hi/english/sci/tech/newsid_1348000/1348820.stm

Warrender C., Forrest S. and Pearlmutter B. (1999), "Detecting Intrusion Using Calls: alternative data models", *Symposium on Security and Privacy*