Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/cose

Computers & Security

Performance evaluation of anomaly-detection algorithms for mouse dynamics

Chao Shen [a], Zhongmin Cai [a,*], Xiaohong Guan [a], Roy Maxion [b]

[a] Xi'an Jiaotong University, No.28 Xianning West Road, Xi'an, 710049, China
[b] Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, 15213, USA

ARTICLE INFO

Article history:
Received 15 April 2014
Accepted 14 May 2014
Available online 27 May 2014

Keywords:
Anomaly detection
Behavioral biometrics
Mouse dynamics
Performance evaluation
Benchmark study
Algorithm comparison

ABSTRACT

Mouse dynamics—the analysis of mouse operating behaviors to identify users—has been proposed for detecting impostors. Since many anomaly-detection algorithms have been proposed for this task, it is natural to ask how well these algorithms perform and how they compare with each other (e.g., to identify promising research directions). This paper presents a performance-evaluation study of a range of anomaly-detection algorithms in mouse dynamics on an equal basis. We collected a mouse-dynamics data set consisting of 17,400 samples from 58 subjects, developed a repeatable evaluation methodology, and implemented and evaluated 17 detectors from the mouse-dynamics and pattern-recognition literatures. Performance is measured in terms of detection accuracy, sensitivity to training sample size, usability with respect to sample length, and scalability with respect to the number of users (user space). The six top-performing detectors achieve equal-error rates between 8.81% and 11.63% with a detection time of 6.1 s; detector performance improves as training sample size and sample length increase and becomes saturated gradually; detector performance decreases as user space becomes large, but only small fluctuations with the error range are apparent when the space size exceeds a certain number. Along with the shared data and evaluation methodology, the results constitute a benchmark for comparing detectors and measuring progress.

1. Introduction

Attackers masquerading as legitimate users have always been a serious concern in the cyber-security setting. Threats from these attackers, who misuse the fabricated or leaked identity credentials for malicious purposes, have overtaken malware and viruses as the most reported security incidents, according to a recent report from the US Computer Security Institute (CSI) (Richardson, 2010). The use of authentication mechanisms is the most common approach to address this problem. But the widespread use of conventional password methods may not offer us enough confidence that our "information gold" will be well protected, given the weaknesses of password being easily stolen and forgotten. Recent events, such as large-scale password leakages (Velazco, 2012; Waugh, 2014), have also brought this concern to the frontline of attention.

One emerging approach that has shown growing interest within the research community (e.g., the Active Authentication program sponsored by DARPA (DARPA, 2012)) is mouse dynamics analyzing users' mouse operating behavior as a behavioral biometric to discriminate among users. Compared with other biometrics, such as fingerprints and iris scan (Jain et al., 2006), mouse dynamics has the advantage of being

* Corresponding author.
E-mail addresses: cshen@sei.xjtu.edu.cn, shenchao.cmu@gmail.com (C. Shen), zmcai@sei.xjtu.edu.cn (Z. Cai), xhguan@sei.xjtu.edu.cn (X. Guan), maxion@cs.cmu.edu (R. Maxion).
http://dx.doi.org/10.1016/j.cose.2014.05.002
0167-4048/

unobtrusive and requiring no specialized hardware to acquire biometric data. When trying to get access to a computer, this biometric usually requires the user to provide a login name and to perform a certain sequence of mouse operations. The features extracted from mouse operations are compared to the profile of a legitimate user. The user would be authenticated and accepted if a match occurs; otherwise the access would be rejected. Furthermore, the user's mouse operations could be continuously analyzed in subsequent interactions to enable active and continuous authentication.

Mouse dynamics has become an active research topic over the last decade. Many anomaly-detection algorithms have been proposed to detect impostors using this biometric feature (see the review in Section 2). It is natural, therefore, to ask how well these detectors perform and how they compare with each other. The primary reason to assess and compare the detectors is to evaluate whether a detector is sufficiently reliable to be put into practice. Since mouse dynamics is still in its early age, and has not reached an acceptable level of performance (e.g., the European standard (CENELEC, 2002)), a secondary reason is to drive progress toward better results. These results would provide the research community with useful insight into what detector characteristics can reduce the error rates, investigating the flexibility and generalization capability of the detectors, and identifying promising directions for further improvement of this biometric.

However, no earlier work has been done on evaluating and measuring the performance of various anomaly detectors. The reasons may be the following: 1) unavailability of a common data set, and 2) inconsistent evaluation methodologies across studies. Most existing approaches have reported good authentication results on limited-size data sets, but these results are hard to compare to each other. Too many factors vary from one study to another. Specifically, most approaches in the literature (1) used different data sets; (2) collected data in different environments; (3) employed different detectors; and (4) used different evaluation procedures. Additionally, known theoretical results for error bounds and probabilities of most detectors are often based on ideal distributions of data conditional densities and/or on infinite samples. Thus a purely theoretical evaluation of a wide variety of detectors for finite samples seems to be difficult and less useful. Additionally, if we are to evaluate the state of the art in mouse dynamics and to measure the future progress, we need a shared benchmark data set and a repeatable evaluation procedure to conduct comparisons, because only then can the error rates of detectors be properly measured and compared.

In this study, our objective is to present a performance evaluation of a range of detectors in mouse dynamic analysis on an equal basis. To achieve this goal, we establish a public data set, develop a repeatable evaluation methodology, and measure a range of mouse-dynamics detectors from various perspectives. The main contributions are summarized as follows:

- A public benchmark mouse-dynamics data set is established, not only for this study, but also to foster future research. This data set contains 17,400 mouse-behavior data samples from 58 subjects. To the best of our knowledge, our group is the first to publish a shared mouse-dynamics data set in this field.
- A repeatable and objective evaluation procedure is developed to evaluate a range of detectors for mouse dynamics analysis. Since the anomaly detectors were evaluated using the same data, under the same conditions, and using the same procedure, it is possible to attribute differences in performance to the detectors and not to different experimental conditions.
- We conduct a rigorous comparison of 17 detectors from the mouse-dynamics and pattern-recognition literatures. We establish which detectors have top detection accuracies using the Wilcoxon signed-rank test and Hochberg's procedure, and we identify promising research directions in mouse-dynamics detectors. This enables us to discover shared detector characteristics and promising strategies among the detectors for reducing error rates.
- We examine the detectors in terms of their sensitivity to training sample size, usability with respect to sample length, and scalability with respect to the number of users, to further examine the flexibility and generalization capability of these detectors in mouse dynamics. We also discuss a number of avenues for future research.
- To our knowledge, this study is the first to systematically evaluate and compare anomaly-detection algorithms in mouse dynamics. Our results, in combination with the shared data set and repeatable evaluation methodology, constitute a benchmark for comparing detectors and measuring progress in mouse dynamics.

The remainder of this paper is organized as follows: Section 2 provides background, and reviews anomaly-detection algorithms in mouse dynamics. Section 3 describes the experimental data set and the feature-extraction method. Section 4 introduces the evaluated anomaly detectors. Section 5 develops the evaluation methodology. Section 6 shows evaluation results in terms of detection accuracy, sensitivity to training sample size, usability with respect to sample length, and scalability with respect to the number of users. Section 7 provides a discussion and possible extensions. Section 8 concludes.

2. Background and related work

In this section, we briefly introduce some background on mouse dynamics research, and review various usages for mouse dynamics (e.g., static analysis vs. continuous analysis) and various analysis techniques (e.g., one-class anomaly detection vs. multi-class classification). We then focus on the technique of using anomaly detectors to study mouse behavior for authentication, and we explain why it is so hard to compare different evaluation results in the literature.

2.1. *Review of mouse dynamics*

Similar to keystroke dynamics (Hosseinzadeh and Krishnan, 2008; Obaidat and Sadoun, 1997; Yong et al., 2005; Killourhy and Maxion, 2009; Maxion and Killourhy, 2010), mouse dynamics, refers to a newly emerging way of recording