Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Science of Computer Programming 60 (2006) 1–26

Science of Computer Programming

www.elsevier.com/locate/scico

Probabilistic anomaly detection in distributed computer networks

Mark Burgess

Oslo University College, Cort Adelers gate 30, 0254 Oslo, Norway

Received 6 July 2004; received in revised form 3 April 2005; accepted 8 June 2005
Available online 27 July 2005

Abstract

Distributed host-based anomaly detection has not yet proven practical due to the excessive computational overhead during training and detection. This paper considers an efficient algorithm for detecting resource anomalies in event streams with either Poisson or long tailed arrival processes. A form of distributed, lazy evaluation is presented, which uses a model for human–computer interaction based on two-dimensional time and a geometrically declining memory to yield orders of magnitude improvements in memory requirements. A three-tiered probabilistic method of classifying anomalous behaviour is discussed. This leads to a computationally and memory economic means of finding probable faults amongst the symptoms of network and system behaviour.

Keywords: Machine learning; Anomaly detection; Data-mining

1. Introduction

Computer anomaly detection is about discerning regular and irregular patterns of behaviour, in the variables that characterize computer systems. The detection of anomalies in computer systems has often been pursued as the unambiguous goal of searching for potential breaches of security; it often goes hand in hand with Network Intrusion Detection, in which content analyses of data are performed in real time with the aim of revealing

E-mail address: Mark.Burgess@iu.hio.no.

0167-6423/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.scico.2005.06.001

suspicious activity [18,42,25,29,34]. However, this is only one application for anomaly detection; computers can also be approached as self-regulating systems that respond to changes in their environment in order to stabilize their own behaviour. In that case, anomaly detection becomes an integral part of the system's regulatory process. Previously, the cost of performing such an analysis on every host has been prohibitive, but this paper will suggest a way of overcoming this difficulty.

Anomaly detectors apply machine learning and analysis to see whether any long term trends can be found in data. One such approach was suggested in the early 1990s and has recently been revived [30,21]. Automated self-regulation in host management has also been discussed in Refs. [7,9,8], as well as adaptive behaviour [51] and network intrusion detection [46,29]. Other authors have likened such mechanisms to immune systems, striking the analogy between computers and other collective systems in sociology and biology [33,24,8].

The ultimate aim of anomaly detection systems is to have adaptive behaviour that responds in 'real time', so that problematical events can be countered as quickly as possible. However, normal behaviour can only be determined by learning about past events: trends take time to learn and analyse. This paradox can only be resolved by modelling future behaviour, on the basis of a statistical idealization of the past and an observation of the present (like weather forecasting). Even then, a timely response requires a rapid processing of observations. The computational burden of real-time anomaly detection can be considerable. One would thus like to spread the burden as far as possible around the network to minimize the load at any particular place.

This paper is motivated by two goals: to develop an efficient method of anomaly detection that avoids bottlenecks, and implements 'lazy evaluation' to avoid unnecessary computational burden; and to develop a language for expressing one's *policy* about what constitutes an anomalous occurrence, relative to what has already been learned about the signal in the past. We shall make some progress towards both of these goals. The paper is organized as follows:

(1) We begin with a brief summary of the idea of host-based anomaly detection, its aims and motivations in relation to the future challenges of mobile and pervasive computing.
(2) Existing techniques for mapping out empirical data characteristics are summarized and appropriate statistical measures for discussing normality are identified.
(3) The notion of policy is then introduced, to account for the arbitrary aspects of data analysis, such as threshold values and the representation of corroborating environmental information that is not represented in the learning abilities of the nodes.
(4) On the basis of the known characteristics of host data, a pseudo-periodic parametrization of time series is developed, which partitions the arrival process into weekly units. Some comments are made about data distributions and the implications for machine learning.
(5) A description of the limited span, unsupervised learning algorithm, with predictable 'forgetting power', is presented.
(6) Finally, a multi-stage classification of data is proposed, where a response is instigated only if a probabilistic detector signals a *probably significant* event (lazy evaluation).