Simulation Modelling Practice and Theory 82 (2018) 103–115

Contents lists available at ScienceDirect

Simulation Modelling Practice and Theory

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

A hierarchical hybrid framework for modelling anomalous behaviours

Fabrizio Angiulli[a], Luciano Argento[a], Angelo Furfaro[a,b,*], Andrea Parise[b]

[a] Dipartimento di Ingegneria Informatica, Modellistica, Elettronica e Sistemistica, Università della Calabria, P. Bucci 41C Rende (CS) 87036, Italy

[b] Open Knowledge Technologies s.r.l. Piazza Vermicelli, Rende (CS) 87036, Italy

ARTICLE INFO

Article history:
Received 28 November 2017
Revised 18 December 2017
Accepted 20 December 2017

Keywords:
Software framework
Anomalous behaviour modelling
Anomaly detection
Signature detection
Data analysis

ABSTRACT

The presence of anomalies in collected information, i.e. data that deviates substantially from what is normally expected, is a valuable source of knowledge and its discovery has many practical applications. Anomaly-detection approaches rely on building models that suitably describe data patterns deemed as normal, however they may incur the generation of a considerable amount of false positives. Signature-based techniques, which exploit a prior knowledge base of anomalous patterns, are able to effectively detect them but fail in identifying anomalies which did not occur previously. Hybrid anomaly detection systems combine the two approaches in order to obtain better detection performances. This paper proposes a framework, called *HALF*, that allows to develop hybrid systems by combining available techniques, coming from both approaches. *HALF* is able to operate on any data type and provides native support to online learning, or concept drifting. This enables the incremental updating of the knowledge bases used by the techniques. *HALF* has been designed to accommodate multiple mining algorithms by organizing them in a hierarchical structure in order to offer an higher and flexible detection capability. The framework effectiveness is demonstrated through two case studies concerning a network intrusion detection system and a steganography hunting system.

1. Introduction

Anomaly detection refers to the problem of finding patterns in data that do not conform to what it is expected [1], that are mainly known as anomalies or outliers in different application domains. Anomaly detection is a common problem to many areas such as fraud detection [2], speech recognition [3], military surveillance for enemy activities [4], intrusion detection for cybersecurity [5] and detection of anomalies in astronomical data [6,7]. This wide range of applications is due to the fact that very often anomalies are sources of significant or critical knowledge that can be derived from raw data. Examples of anomalies are banking transactions from unusual places which could indicate credit card or identity theft, in the context of fraud detection; presence of malformed strings in network packets, for cyber security; anomalous energy consumption, both for cybersecurity [8] and energy consumption [9], and so on.

* Corresponding author: Dipartimento di Ingegneria Informatica, Modellistica, Elettronica e Sistemistica, Università della Calabria, P. Bucci 41C Rende (CS) 87036, Italy.

E-mail addresses: f.angiulli@dimes.unical.it (F. Angiulli), l.argento@dimes.unical.it (L. Argento), a.furfaro@dimes.unical.it (A. Furfaro), andrea.parise@okt-srl.com (A. Parise).

https://doi.org/10.1016/j.simpat.2017.12.013
1569-190X/

A data analyst typically focuses on a particular class of anomalies for his study, therefore not all anomalies are always interesting to find.

Given a specific class of anomalies, regardless of the domain considered, two types of anomalies can be distinguished: known and unknown. The former are represented by patterns, analyzed and documented extensively and learned from particular events of interest. These patterns can be described by means of rules or signatures. For example, in the field of cybersecurity such anomalies are symptoms of known attacks, while in the medical field they could consist of chromosomal anomalies that suggest the presence of a disease [10]. Unknown anomalies are patterns that were not yet observed, or for which there is poor knowledge. Their nature can be characterized, to some extent, leveraging knowledge coming from historical data and/or abnormal events from which they were generated. In the field of cybersecurity, examples of unknown anomalies are zero-day attacks, i.e. attempts to exploit an either unknown or not yet published vulnerability. As another example in astronomy, an abnormal refraction of light detected by a telescope may suggest the presence of a not yet identified celestial body.

The problem of anomaly detection has received a lot of attention since the beginning of the 19th century in the statistics community [11]. Recently, interesting advances in the application of machine learning algorithms for anomaly detection have been achieved. Over the years the research community has proposed a considerable variety of techniques, many of which are very specific to their application domains [12].

Signature-based techniques aim at analyzing the data to find one or more matches with a set of rules or signatures. Signatures can include specific strings or regular expressions that characterize one or more classes of anomalies. These approach typically generates few false positives, however they fails to detect unknown anomalies or variants of known ones.

Anomaly-based techniques use suitable models that represent normal data and classify as anomalous data which deviate considerably from what expected by the model. Unlike signature-based techniques, they are able to discover both known and unknown anomalies. The main problem of these approaches is the generation of a not negligible amount of false positives. Despite both classes of techniques have been widely used, the above issues have not yet been fully resolved. Some researchers have thus focused their attention on hybrid techniques, which are considered a viable solution. Such systems are designed to include the features of both signature-based and anomaly-based techniques in the attempt to gain the advantages of both of them and, at the same time, to mitigate their defects.

Another important issue, which is very relevant in some contexts, regards the fact that data may change over time thus leading to a progressive degradation of the accuracy of the models employed by the anomaly detection techniques. This phenomenon is known as *concept drifting* [13]. There are three types of variations that may occur in the data source, which have been described in literature: *sudden drift, recursive drift* and *gradual drift*. Sudden drift consists of an instantaneous change of concept, recursive drift is an alternation of two or more concepts over time with or without a periodical order and gradual drift denotes a smooth changing of the concept. In order to learn and maintain an accurate detection level in presence of concept drift it is fundamental to detect its occurrence and suitably adapt the model to classify new data. Techniques developed to overcome concept drift can be divided into three categories [14]: adaptive based, learners which modify the training set, ensemble techniques.

In the literature, there are many different approaches to the identification of anomalies, e.g. relying on statistics, machine learning, data mining. Each devised technique has unique features, strengths and drawbacks. Combining more of them together can allow to achieve better results.

This paper proposes a flexible multi-domain framework, called *HALF*, that generalizes the problem of anomaly detection. The framework is designed to embrace both signature-based and anomaly-based techniques. In addition, it makes possible to combine the use of different models for the analysis of data, organized in a hierarchical structure. Given its nature, *HALF* can work on any kind of data, unlike the existing works which are bound to specific fields of application. Furthermore, one of the most important features of the framework is the ability to natively support data evolution, i.e. concept drift, regardless of the presence or not of this capability in the techniques employed for the analysis. In this way, even if a technique is designed to work on static data, it becomes fully usable also on dynamic data.

The rest of the paper is organized as follows. Section 2 summarizes the related work. Section 3 details the *HALF* architecture while Section 4 explains the design choices. Section 5 presents two case-studies and finally Section 6 concludes the paper.

2. Related work

Hybrid approaches, subject of great interest in the research community, aim at exploiting the strengths of individual components, to obtain benefits from their combination. They are widely used in the domain of cyber security, especially for monitoring network activities, where many hybrid intrusion detection systems have been proposed [15].

A hybrid network intrusion detection system (HNIDS), which combines two anomaly techniques, i.e. packet header anomaly detection (*PHAD*) [16] and network traffic anomaly detection (*NETAD*) [17], with a misuse technique, was proposed in [18]. In particular, by using Snort [19] as the misuse engine, they exploited its pre-processors in order to integrate both *PHAD* and *NETAD*. Packets are first analyzed in sequence by *PHAD*, then by *NETAD* and at last by Snort. The reported results show the effectiveness of this hybrid approach.