Available online at www.sciencedirect.com

ScienceDirect

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

Computers & Security

PCA-based multivariate statistical network monitoring for anomaly detection

José Camacho *, *Alejandro Pérez-Villegas*, *Pedro García-Teodoro*, *Gabriel Maciá-Fernández*

Department of Signal Theory, Telematics and Communications, School of Computer Science and Telecommunications – CITIC, University of Granada, Granada, Spain

ARTICLE INFO

Article history:
Received 14 September 2015
Received in revised form 18 February 2016
Accepted 18 February 2016
Available online 2 March 2016

Keywords:
Multivariate statistical process control
Network monitoring
Network security
Principal component analysis
Anomaly detection

ABSTRACT

The multivariate approach based on Principal Component Analysis (PCA) for anomaly detection received a lot of attention from the networking community one decade ago, mainly thanks to the work of Lakhina and co-workers. However, this work was criticized by several authors who claimed a number of limitations of the approach. Neither the original proposal nor the critic publications were completely aware of the established methodology for PCA anomaly detection, which by that time had been developed for more than three decades in the area of industrial monitoring and chemometrics as part of the Multivariate Statistical Process Control (MSPC) theory. In this paper, the main steps of the MSPC approach based on PCA are introduced; related networking literature is reviewed, highlighting some differences with MSPC and drawbacks in their approaches; and specificities and challenges in the application of MSPC to networking are analyzed. All of this is demonstrated through illustrative experimentation that supports our discussion and reasoning.

1. Introduction

The outstanding capability of multivariate analysis to detect anomalies has been recognized in several domains, including industrial monitoring (Camacho et al., 2009; Chen et al., 2002; Hu et al., 2008; Nomikos and MacGregor, 1994) and networking (Brauckhoff et al., 2009; Chatzigiannakis and Androulidakis, 2009; Lakhina et al., 2004; Münz, 2010; Ringberg et al., 2007). The use of multivariate analysis for anomaly detection is typically referred to as Multivariate Statistical Process Control (MSPC) (Ferrer, 2014). A main tool in MSPC is Principal Component Analysis (PCA).

The pioneering work by Lakhina et al. (2004) introduced the use of PCA for network anomaly detection. Their approach received a lot of attention from the networking community one decade ago, and thus a variety of other proposals has been developed based on it. However, the approach was also criticized by a number of papers. Ringberg et al. (2007) claimed that it is sensitive to calibration settings. In particular, that:

1. The false positive rate is very sensitive to small differences in the number of principal components in the normal subspace.
2. The effectiveness of PCA is sensitive to the level of aggregation of the traffic measurements.

* *Corresponding author.*
E-mail address: josecamacho@ugr.es (J. Camacho).
http://dx.doi.org/10.1016/j.cose.2016.02.008

3. A large anomaly may inadvertently pollute the normal subspace, and go undetected.
4. Correct diagnosis is an inherently challenging problem.

Here we argue that these supposed problems are the result of flaws in adopting PCA to the anomaly detection field. It should be noted that such flaws are found not only in the original approach but also in its detractors. Although Lakhina et al. noted that similar approaches to theirs were already developed in the chemical engineering area, the bulk of the (by that time) well-established theory of MSPC based on PCA was ignored in their papers.

In this work, we review the theory of PCA-based MSPC, highlighting (or stressing) the differences with Lakhina et al. and posterior approaches and ilustrating this differences with examples. We refer to the approach that follows the MSPC theory for anomaly detection in communication networks as the Multivariate Statistical Network Monitoring (MSNM). The last term in MSNM, "monitoring", has been preferred to control, which is seldom used in the networking community. Furthermore, the term "control" has a different meaning in fields other than statistics, such as automatic feedback control (Camacho, 2007; MacGregor and Kourti, 1995).

The rest of the paper is organized as follows. Section 2 reviews the principal works on PCA-based network anomaly detection. Section 3 presents fundamentals on statistical process control, in particular on the use of PCA-based MSPC. After that, Section 4 discusses the necessary pre-processing for networking data to be analyzed with PCA, while the proper processing for dynamic modeling is subsequently described in Section 5. The discussion and argumentation carried out until this point are demonstrated by means of some illustrative examples in Section 6. Finally, Section 7 summarizes the main contributions of the work and future challenges.

2. Related work

Supervising computer and network systems is a key topic in the literature from several decades ago. The main purpose of existent solutions in the filed is the early detection of potential failures and malfunctions. From this, some recovery actions could be taken in order to restore the normal desired operation for the monitored environment.

The terms "failure" and "malfunction" must be interpreted as a global concept that can be caused by a number of different reasons, either accidental or not. One of the most studied causes is the one deliberately carried out by malicious users through attacks aimed at exploiting some system vulnerability. Whichever the origin of the failure or malfunction, the goal of the monitoring and detection systems is similar: to prevent the environment from decreasing its performance or even from crashing. For that, the usual procedure when determining the occurrence of some kind of "anomaly" (*i.e.*, a certain deviation from the normal expected behavior of the monitored environment) is to trigger an alarm as a previous step to solve the problem.

Among several other existent anomaly detection paradigms (Bhuyan et al., 2014; Garcia-Teodoro et al., 2009), statistical solutions have been widely adopted (Om and Hazra, 2012). In particular, multivariate approaches such as PCA were adopted several years ago (Bodenham and Adams, 2013; Kanaoka and Okamoto, 2003; Qu et al., 2005; Shyu et al., 2003), their unsupervised nature being the main benefit argued in comparison with other solutions. As previously stated, maybe the most referred work is that of Lakhina et al. (2004). Based on it, several further proposals have been developed in the literature.

Authors in Huang et al. (2006) introduce a network anomaly detection for large distributed systems. It is based on a stochastic matrix perturbation analysis that characterizes the trade-off between the accuracy of anomaly detection and the amount of data communicated over the network. On the other hand, Kwitt and Hofmann (2007) discuss the problem of contaminated training data and propose to use PCA on the basis of robust estimators to overcome the necessity of a supervised preprocessing step for anomaly detection in the context of intrusion detection systems. Also Hakami et al. (2008) and Rubinstein et al. (2008) highlight the advantage of PCA in avoiding the need of labeled training datasets in comparison with other detection schemes.

Kim et al. (2009) present a higher-order singular value decomposition (HOSVD) and higher-order orthogonal iteration (HOOI) algorithms on network traffic anomaly detection by rearranging the data in tensor formats. Simulation results show that the higher-order methods improve the detection performance while also reduce the complexity for large-scale networks. The work in Liu et al. (2010) tries to solve scalability problems of PCA. For that, a sketch-based streaming PCA algorithm for the network-wide traffic anomaly detection in a distributed fashion is proposed.

Authors in Magan-Carrion et al. (2015) introduce a PCA-based methodology to detect anomalies related to potential losses of data in WSNs. Based on this, a subsequent data recovery procedure is also contributed. This relies on the exploitation of the spatial correlation inherent in WSNs. Different routing strategies to collect all the information around the network are analyzed to evaluate the suitability of the approach.

Reference Livani and Abadi (2010) uses distributed principal component analysis (DPCA) and fixed-width clustering (FWC) in order to establish a global normal profile and to detect anomalies. The process of establishing the global normal profile is distributed among all sensor nodes. Authors also use weighted coefficients and a forgetting curve to periodically update the established normal profile. A similar work in obtaining user profiles in communication environments is that in Dusi et al. (2012).

In Callegari et al. (2011), the "classical" PCA approach is complemented with the Kullback–Leibler divergence to improve detection results. Similarly, Xie et al. (2011) combines PCA with distance-based anomaly detection (DB-AD) to reduce dimensionality. Authors in Novakov et al. (2013) combine PCA with wavelet algorithms for network traffic anomaly detection. References Delimargas et al. (2014) and Liu et al. (2014) study PCA variants to solve the calibration sensitivity. Like the latter, Kanda et al. (2013) uses a entropy-based PCA.

As aforementioned, almost all of the existent works on PCA-based anomaly detection in networking are developed taking as a base the work by Lakhina et al. (2004). This way, all of them present similar disadvantages. The main points in which the