Information Sciences 178 (2008) 3840–3859

Contents lists available at ScienceDirect

Information Sciences

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

Local anomaly detection for mobile network monitoring

Pekka Kumpulainen [a,*], Kimmo Hätönen [b,1]

[a] Tampere University of Technology, Automation Science and Engineering, Box 692, 33101 Tampere, Finland
[b] Nokia Siemens Networks, Research, Technology and Platforms, Box 6, 02022 Nokia Siemens Networks, Finland

ARTICLE INFO

Article history:
Received 2 November 2007
Received in revised form 25 April 2008
Accepted 29 May 2008

Keywords:
Local anomaly detection
Outlier
Mobile networks
System log
Self-organizing map
Adaptive thresholds

ABSTRACT

Huge amounts of operation data are constantly collected from various parts of communication networks. These data include measurements from the radio connections and system logs from servers. System operators and developers need robust, easy to use decision support tools based on these data. One of their key applications is to detect anomalous phenomena of the network. In this paper we present an anomaly detection method that describes the normal states of the system with a self-organizing map (SOM) identified from the data. Large deviation in the data samples from the SOM nodes is detected as anomalous behavior. Large deviation has traditionally been detected using global thresholds. If variation of the data occurs in separate parts of the data space, the global thresholds either fail to reveal anomalies or reveal false anomalies. Instead of one global threshold, we can use local thresholds, which depend on the local variation of the data. We also present a method to find an adaptive threshold using the distribution of the deviations. Our anomaly detection method can be used both in exploration of history data or comparison of unforeseen data against a data model derived from history data. It is applicable to wide range of processes that produce multivariate data. In this paper we present examples of this method applied to server log data and radio interface data from mobile networks.

1. Introduction

Detection of anomalies or outliers is an important task in data analysis. Locating rare or exceptional parts of the data can reveal new valuable information from the system. As Kruskal wrote in 1960 [18]: *An apparently wild* (*or otherwise anomalous*) *observation is a signal that says*: "*Here is something from which we may learn a lesson*, *perhaps of a kind not anticipated beforehand*, *and perhaps more important than the main object of the study*".

Mobile telecommunication networks are complex distributed systems. The amount of data traffic and number of services is ever-growing. Network monitoring and management tools collect and use large amounts of data provided by network elements. This data includes, for example, performance measurements from the radio interface and log data from application servers. A data set collected from an operational GSM network during one day consists of several gigabytes of data. It is impossible for network operators to analyze all the data manually, especially in multivariate space, where it is impossible to visualize all the dimensions of the data space simultaneously. Therefore, automated multivariate methods are needed to analyze the data sets.

Fault and intrusion detection and network optimization are examples of continuously ongoing network monitoring tasks. An operational GSM network can consist of thousands of base stations, hundreds of base station controllers and dozens of management system servers. Monitoring tasks can be divided to several subtasks focusing on geographical part of the

* Corresponding author. Tel.: +358 40 8490930; fax: +358 3 31152171.
E-mail addresses: pekka.kumpulainen@tut.fi (P. Kumpulainen), kimmo.hatonen@nsn.com (K. Hätönen).
[1] Senior specialist: security research.

0020-0255/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2008.05.038

network or a group of services. The task execution is based on the selected set of data, which can contain millions of events or several hundreds of time series, each of which is representing a history of one performance indicator of a single network element or process. To find malfunctions or sub-optimally behaving network elements or processes, operators use tools that analyze these time series and typically search for data samples whose values exceed pre-defined thresholds or their combinations [21]. There can be several hundreds of thresholds stored and maintained in analysis tools.

Universally applicable automated analysis tools are practically impossible to create due to intricacy of the systems and diverse information requirements of the users. One suitable task for automatic data mining tools is to model the normal or most common behavior of the system and to detect unusual situations. The data collected from various parts of mobile network usually include samples from normal states as well as abnormal situations. With such data we have to assume that the vast majority of the data is from normal functionality and the rare states present some sorts of anomalies or outliers. Thus events outside the common model can be regarded as exceptional and thus providing new information about the behavior of the system under study.

Anomalies or outliers in data are often signs of malfunction or otherwise undesired performance and, thus, they should be detected as soon as possible. In multivariate data there is a vast variety of sources causing anomalous behavior. Therefore a human expert is usually required to further analyze these situations. The purpose of the analysis tools is to filter out the majority of the data and to present the user only a limited amount of data, containing the most useful new information of the system.

A general definition for an outlier was given by Hawkins [9]: *An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism*. This definition is very extensive but it gives no guidelines how to determine whether an observation is an outlier or not.

Various statistical methods have been used in outlier detection [1] and for online monitoring purposes there are specific tests in statistical process control (SPC) [8]. These are mostly univariate methods and rely on the knowledge of the underlying distribution. Multivariate SPC methods [7] also assume multinormal distributions. However, these methods are not well applicable in telecommunication network management, since neither radio performance measurements, nor log activity counters usually follow any known distribution. Some traffic related features have heavy tail distributions and are closely related to Ethernet traffic [38], which is self-similar by nature [22]. Poisson models, for example do not fit the network traffic [28] and more complicated models are required, such as mixtures of exponentials [6]. The variables used in network management are aggregated from several counters. The original counters and the formulas to calculate the KPI (Key Performance Indicator) are often company confidential. We encounter a variety of distributions both in server log activity and radio interface performance measurements. Examples of distributions are depicted in Fig. 1. From left to right we have two histograms presenting aggregated log activity variables and the last one is an example of a radio interface performance KPI. These examples include skewed heavy tailed and multimode distributions and the real data contains a variety of other types of distributions. In practice it is impossible to use any single distribution model and a mixture of symmetric distributions like GMM (Gaussian Mixture Model) do not fit these data very well. Therefore we prefer a method which does not need any assumptions about the underlying distribution.

ICA (Independent Component Analysis) has been applied for Multivariate SPC in order to perform better with non-normally distributed variables [15]. Knorr et al. present the notion of distance based outliers [16]. This requires no assumptions of distribution, but is based on the distances between the data samples. Principal component classifiers have also been used in anomaly detection [32].

All the methods mentioned above are global in a sense that they treat all the data set as one group. If the data are clustered, they may fail. Definition of local outliers by Breunig et al. [3] takes the clustering structure of the data into account.

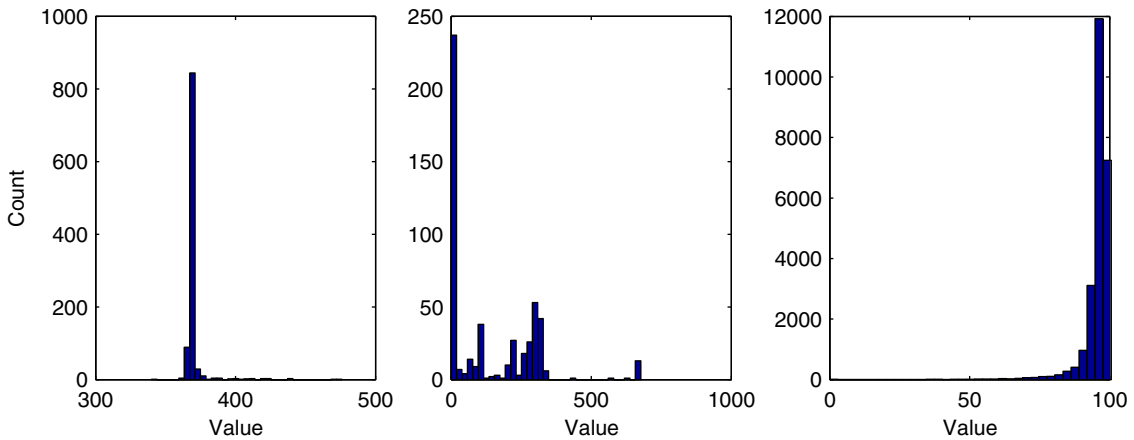

Fig. 1. Histograms of three types of variables. The ones on the left and in the middle are log activity counters and the one on the right is an example of a performance KPI from a radio interface.