Automatica 50 (2014) 1459–1465

Contents lists available at ScienceDirect

Automatica

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

Brief paper

Scalable anomaly detection in large homogeneous populations☆

Henrik Ohlsson [a,b], Tianshi Chen [a,1], Sina Khoshfetrat Pakazad [a], Lennart Ljung [a], S. Shankar Sastry [b]

[a] Department of Electrical Engineering, Linköping University, SE-581 83 Linköping, Sweden
[b] Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, CA, USA

ARTICLE INFO

Article history:
Received 3 October 2012
Received in revised form
21 December 2013
Accepted 19 February 2014
Available online 28 March 2014

Keywords:
Anomaly detection
Outlier detection
Multi-hypothesis testing
Distributed optimization
System identification

ABSTRACT

Anomaly detection in large populations is a challenging but highly relevant problem. It is essentially a multi-hypothesis problem, with a hypothesis for every division of the systems into normal and anomalous systems. The number of hypothesis grows rapidly with the number of systems and approximate solutions become a necessity for any problem of practical interest. In this paper we take an optimization approach to this multi-hypothesis problem. It is first shown to be equivalent to a non-convex combinatorial optimization problem and then is relaxed to a convex optimization problem that can be solved distributively on the systems and that stays computationally tractable as the number of systems increase. An interesting property of the proposed method is that it can under certain conditions be shown to give exactly the same result as the combinatorial multi-hypothesis problem and the relaxation is hence tight.

1. Introduction

In this paper we study the following problem: we are given N systems and we suspect that $k \ll N$ of them behave differently from the majority. We do not know beforehand what the normal behavior is, and we do not know which k systems behave differently. This problem is known as an *anomaly detection problem* and has been discussed *e.g.,* in Chandola, Banerjee, and Kumar (2009); Chu, Gorinevsky, and Boyd (2011); Gorinevsky, Matthews, and Martin (2012). It clearly has links to *change detection* (*e.g.,* Basseville & Nikiforov, 1993, Gustafsson, 2001 and Patton, Frank, & Clark, 1989) but is different because the detection of anomalies is done by comparing systems rather than looking for changes over time.

The anomaly detection problem typically becomes very computationally demanding, and it is therefore of interest to study *distributed solutions*. A distributed solution is also motivated by that many anomaly detection problems are spatially distributed and lack a central computational unit.

Example 1 (*Aircraft Anomaly Detection*)**.** In this example we consider the problem of detecting abnormally behaving airplanes in a large homogeneous fleet of aircrafts. Homogeneous here means that the normal aircrafts have similar dynamics. This is a very relevant problem (Chu et al., 2011; Gorinevsky et al., 2012) and of highest interest for safety in aeronautics. In fact, airplanes are constantly gathering data and being monitored for this exact reason. In particular, so called *flight operations quality assurance* (FOQA) data are collected by several airlines and used to improve their fleet's safety.

As showed in Chu et al. (2011), faults in the angle-of-attack channel can be detected by studying the relation between the angle of attack, the dynamic pressure, mass variation, the stabilizer deflection angle, and the elevator deflection. The number of airplanes in a fleet might be of the order of hundreds and data from a couple of thousand flights might be available (200 airplanes and data from 5000 flights were used in Chu et al., 2011). Say that our goal is to find the 3 airplanes among 200 airplanes that are the most likely to be anomalous to narrow the airplanes that need manual inspection. Then, we would have to evaluate roughly 1.3×10^6 hypothesis (the number of unordered selections of 3 out of 200 airplanes). For each hypothesis, the likelihood for the observed data would then be maximized with respect to the unknown parameters and the most likely hypothesis accepted. This is clearly a very computationally challenging problem.

☆ The material in this paper was partially presented at the 16th IFAC Symposium on System Identification (SYSID 2012), July 11–13, 2012, Brussels, Belgium. This paper was recommended for publication in revised form by Associate Editor Michele Basseville under the direction of Editor Torsten Söderström.

E-mail addresses: ohlsson@isy.liu.se (H. Ohlsson), tschen@isy.liu.se (T. Chen), sina.kh.pa@isy.liu.se (S.K. Pakazad), ljung@isy.liu.se (L. Ljung), sastry@eecs.berkeley.edu (S. Shankar Sastry).

[1] Tel.: +46 0 13 28 22 26; fax: +46 0 13 28 26 22.

http://dx.doi.org/10.1016/j.automatica.2014.03.008
0005-1098/

Example 1 considers anomaly detection in a large homogeneous population and is the type of problem we are interested in solving in this paper. The problem has previously been approached using *model based anomaly detection* methods, see *e.g.,* Chandola et al. (2009); Chu et al. (2011) and Gorinevsky et al. (2012). This class of anomaly detection methods is suitable to detect anomalies in systems, as opposed to non-model based methods that are more suitable for finding anomalies in data. Model based anomaly detection methods work under the assumption that the dynamics of normal systems is the same, or equivalently, that the population of systems is homogeneous. The normal dynamics is modeled from system observations and most papers assume that an abnormal-free training data set is available for the estimation, see for instance (Abraham & Box, 1979; Abraham & Chuang, 1989; Fox, 1972). Some papers have been presented to relax this assumption. In *e.g.,* Rousseeuw and Leroy (1987), the use of a regression technique robust to anomalies was suggested.

The detection of anomalous systems is in model based anomaly detection done by comparing system observations and model predictions and often done by a statistical test, see *e.g.,* Desforges, Jacob, and Cooper (1998) and Eskin (2000). However, in non-model based anomaly detection, classification based (Duda, Hart, & Stork, 2000; Tan, Steinbach, & Kumar, 2005), clustering based (Jain & Dubes, 1988), nearest neighbor based (Tan et al., 2005, Chapter 2), information theoretic (Arning, Agrawal, & Raghavan, 1996) and spectral methods (Parra, Deco, & Miesbach, 1996) are also common. See Chandola et al. (2009) for a detailed review of anomaly detection methods. Most interesting and similar to the proposed method is the more recent approach taken in Chu et al. (2011) and Gorinevsky et al. (2012). They simultaneously estimate the regression model for the normal dynamics and perform anomaly detection. The method of Chu et al. (2011) is discussed further in the numerical section. There has also been some work on distributed anomaly detection, *e.g.,* Chatzigiannakis, Papavassiliou, Grammatikou, and Maglaris (2006), Chu et al. (2011) and Zimmermann and Mohay (2006).

The main contribution of the paper is a novel distributed, scalable and model based method for anomaly detection in large homogeneous populations. The method is distributed in the sense that the computations can be distributed over the systems in the population or a cluster of computers. It is scalable since the size of the optimization problem solved on each system is independent of the number of systems in the population. This is made possible by a novel formulation of the multi-hypothesis problem as a sparse problem. The method also shows superior performance and is easier to tune than previously proposed model based anomaly detection methods. Lastly, the method does not need a training data set and a regression model of the normal dynamics is estimated at the same time as abnormal systems are detected. This is particularly valuable since often neither a training data set nor a regression model for the normal dynamics is available.

The remainder of the paper is organized as follows. Section 2 states the problem and shows the relation between anomaly detection and multi-hypothesis testing. Section 3 reformulates the multi-hypothesis problem as a sparse optimization problem and Section 4 gives a convex formulation. The convex problem is solved in a distributed manner on the systems and this is discussed in Section 5. We return to Example 1 and compare to the method of Chu et al. (2011) in Section 6. Finally, we conclude the paper in Section 7.

2. Problem statement and formulation

Assume that the population of interest consists of N systems. Think for example of the N airplanes studied in Example 1. Further assume that there is a linear unknown relation describing the relation between measurable quantities of interest (angle of attack, the dynamic pressure, mass variation, the stabilizer deflection angle, and the elevator deflection in Example 1):

$$y_i(t) = \varphi_i^{\mathsf{T}}(t)\theta_{i,0} + e_i(t), \quad i = 1, \ldots, N, \tag{1}$$

where t is the time index, i indexing systems, $y_i(t) \in \mathbb{R}$ and $\varphi_i(t) \in \mathbb{R}^m$ are the measurement and regressor vector at time t, respectively, $\theta_{i,0}$ is the *unknown* model parameter, and $e_i(t) \in \mathbb{R}$ is the measurement noise. For the ith system, $i = 1, \ldots, N$, let $\{(y_i(t), \varphi_i(t))\}_{t=1}^{\Omega}$ denote the collected data set and Ω the number of observations collected on each system. We assume that $e_i(t)$ is white Gaussian distributed with mean zero and some *unknown* variance σ^2 and moreover, independent of $e_j(t)$ for all $i \neq j$. However, log-concave distributed noise could be handled with minor changes.

We will in the following say that the population behaves *normally* and that none of the systems are abnormal if $\theta_{1,0} = \cdots = \theta_{N,0} = \theta_0$. Conversely, if any system has a model parameter deviating from the *nominal* parameter value θ_0, we will consider that system as *abnormal*.

To solve the problem we could argue like this: suppose we have a hypothesis about which k systems are the anomalies. Then we could estimate the nominal parameters θ_0 by least squares from the rest, and estimate individual θ_i for the k anomalies. Since we do not know which systems are the anomalies, we have to do this for all possible hypotheses: choosing k systems from a set of N leads to a total of

$$c(N, k) = N!/\big((N-k)!k!\big) \tag{2}$$

possible hypotheses. To decide which is the most likely hypothesis, we would evaluate the total misfit for all the systems, and choose that combination that gives the smallest total misfit. If we let γ_j be the set of assumed abnormal systems associated with the jth hypothesis $j = 1, \ldots, c(N, k)$, this would be equivalent to solving the non-convex optimization problem

$$\begin{aligned} \underset{j=1,\ldots,c(N,k)}{\text{minimize}} & \sum_{s\in\gamma_j} \min_{\theta_{j,s}} \sum_{t=1,\ldots,\Omega} \|y_s(t) - \varphi_s^{\mathsf{T}}(t)\theta_{j,s}\|^2 \\ & + \min_{\theta_{j,0}} \sum_{s\notin\gamma_j, t=1,\ldots,\Omega} \|y_s(t) - \varphi_s^{\mathsf{T}}(t)\theta_{j,0}\|^2. \end{aligned} \tag{3}$$

Since we assume that all systems have the same noise variance σ^2, this is a formal hypothesis test. If the systems may have different noise levels we would have to estimate these and include proper weighting in (3).

The difficulty is how to solve (3) when the number of systems N is large. As seen in Example 1, even for rather small examples ($k = 3, N = 200$), the number of hypothesis $c(N, k)$ becomes large and solving problem (3) becomes computationally intractable.

3. Sparse optimization formulation

A key observation to be able to solve the anomaly detection problem in a computationally efficient manner is the reformulation of the multi-hypothesis problem (3) as a sparse optimization problem. To do this, first notice that the multi-hypothesis test (3) will find the k systems whose data are most likely to not have been generated from the same model as the remaining $N - k$ systems. Let us say that j^* was the selected hypothesis and denote the parameter of the ith system by θ_i, $i = 1, \ldots, N$. Then $\theta_{i_1} \neq \theta_{i_2}$ for all $i_1, i_2 \in \gamma_{j^*}$ and $\theta_{i_1} = \theta_{i_2}$ for all $i_1, i_2 \in \{1, \ldots, N\}/\gamma_{j^*}$. Note that $N - k$ systems will have identical parameters. An equivalent way of solving the multi-hypothesis problem is therefore to maximize the likelihood under the constraint that $N-k$ systems are identical.