Available online at www.sciencedirect.com

ScienceDirect

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

A coupled classification – Evolutionary optimization model for contamination event detection in water distribution systems

*Nurit Oliker, Avi Ostfeld**

Faculty of Civil and Environmental Engineering, Technion – Israel Institute of Technology, Haifa 32000, Israel

ARTICLE INFO

Article history:
Received 26 July 2013
Received in revised form
25 October 2013
Accepted 26 October 2013
Available online 8 November 2013

Keywords:
Water distribution systems
Water quality
Water security
Event detection
Support vector machine
Sequence analysis

ABSTRACT

This study describes a decision support system, alerts for contamination events in water distribution systems. The developed model comprises a weighted support vector machine (SVM) for the detection of outliers, and a following sequence analysis for the classification of contamination events. The contribution of this study is an improvement of contamination events detection ability and a multi-dimensional analysis of the data, differing from the parallel one-dimensional analysis conducted so far. The multivariate analysis examines the relationships between water quality parameters and detects changes in their mutual patterns. The weights of the SVM model accomplish two goals: blurring the difference between sizes of the two classes' data sets (as there are much more normal/regular than event time measurements), and adhering the time factor attribute by a time decay coefficient, ascribing higher importance to recent observations when classifying a time step measurement. All model parameters were determined by data driven optimization so the calibration of the model was completely autonomic. The model was trained and tested on a real water distribution system (WDS) data set with randomly simulated events superimposed on the original measurements. The model is prominent in its ability to detect events that were only partly expressed in the data (i.e., affecting only some of the measured parameters). The model showed high accuracy and better detection ability as compared to previous modeling attempts of contamination event detection.

1. Introduction

Securing infrastructure is unquestionably an eminent task, being vital for the population welfare. A water distribution system (WDS) is inherently vulnerable as it comprises numerous exposed elements that can be easily infiltrated. The threat of contamination events in a WDS, deliberate, accidental, or natural, raises growing concern worldwide. Many resources have been invested both in academia and industry for the development of contamination warning systems. The two major issues in the application of such system are the placement of sensors in the network and the analysis of the measured data. The problem of sensor placement was widely explored, featuring over 90 studies related to sensor placement optimization (Hart and Murray, 2010). The vast majority of sensor placement models use the assumption of a "perfect

* *Corresponding author*. Tel.: +972 4 8292782; fax: +972 4 8228898.
E-mail address: ostfeld@tx.technion.ac.il (A. Ostfeld).

0043-1354/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.watres.2013.10.060

sensor", meaning that if a sensor measures any concentration of a contaminant, it will detect it promptly. Few studies assumed that sensors will detect any contaminant above some contamination threshold concentration. Virtually the recognition of a contaminant is a complex task. In the overview of warning systems development, there is a gap in the data analysis element and a need in elaborating the models, and respectively, the evaluation of the detection ability.

Some attempts were made to develop sensors suitable for identifying specific pollutants according to their unique properties [e.g., the use of light scattering for the detection of spectral signature (Adams and Mccarty, 2007)]. The large variety of pollutants, however, made it impossible to deal with all, and problematic to focus just on some. In addition, the task of pollutant recognition was revealed as rather complex. Thus, the approach of specific contaminant recognition was fallen from grace, and a more generic approach was adopted. The latter features the use of typically monitored water quality parameters, such as turbidity, electrical conductivity, pH, and residual chlorine concentration, for the detection of abnormal behavior. The premise of this approach is that a contaminant intruding the system will cause some disturbances in the general water quality measurements. According to this, information from online water quality sensors may provide an early indication of a pollution presence in the WDS. The challenge is then to distinguish between normal behavior of the parameters, and changes triggered by contaminants intrusion.

A few studies utilized general water quality measurements for the purpose of contamination event detection. Murray et al. (2010), Perelman et al. (2012), and Arad et al. (2013) developed event detection models. Those studies feature a parallel analysis of the water quality parameters. Some machine learning methods learn the behavior of each parameter time series, and the expected measured value of the next time step is predicted. That way, the models identify deviations from expected behavior and classify measurement outliers. The estimations of all the parameters are integrated to assess the probability of an event occurrence. Guepie et al. (2012) developed a model based on residual chlorine decay. Their premise was that a contaminant in the WDS will consume a significant fraction out of the measured chlorine and this single parameter will provide valuable sufficient data.

The above studies utilized supervised classification methods. Through employing these methods, the classifier learns to distinguish between normal operation and event time measurements according to a given training data set. The lack of real event time measurements requires the use of simulated contamination events for the training and testing of the models. To maintain generality, and in the absence of adequate knowledge, the models apply some random disturbances to the measured data for representing the contaminant effect. Uber et al. (2007) provided guidelines for event simulation based on contaminant reaction kinetics and uncertainty.

The aforementioned studies utilized water quality time series for the purpose of contamination event detection by applying simultaneous processes of univarite analysis. The parallel processes were followed by an integration of the analyses for the assessment of event occurrence probability. A multivariate analysis differs from the studies that were conducted so far by involving the observation and analysis of more than one parameter at a time.

The objectives of this study are: (1) to apply multivariate analysis of the data, which includes the simultaneous analysis of all variables in a multi-dimensional space. The multivariate analysis produces a very different description of the system, revealing phenomena evolving parameters relationships, (2) to develop a complete and independent system whose input is the on-line measurements and its output is a contamination event alert, and (3) to develop a model that requires minimal calibration.

This work suggests a contamination event detection model supplying an autonomic decision support system. The study applies a weighted support vector machine (SVM) classifier for the detection of outlier measurements, and a following sequence analysis for the events classification.

SVM was introduced by Boser et al. (1992). It is a very popular classification method prominent in its high accuracy and the ability to deal with high-dimensional data. The SVM transforms an un-specified learning task into a linear optimization problem. The training data set is used to create a hyperplane which separates the space into two classes.

A decision boundary, situated in both sides of the hyperplane defines the separating area. The objectives that define the hyperplane parameters are: maximization of the separation area, and minimization of the error of misclassified vectors. Naturally, this is a trade-off, since the more errors enabled, the larger the margin will be.

The SVM problem can be defined as:

$$\underset{W,b,\ \ Si\geq 0}{\text{Minimize}}\left[\tfrac{1}{2}\|W\|^2 + C\sum_{i=1}^{n} S_i\right]$$
$$\text{Subject to}: y_i(W^T x_i + b) \geq 1 - S_i \quad i = 1,\ldots,n \tag{1}$$

where W is the normal vector to the hyperplane, C is the classifier parameter, S_i are the slack variables, n is the number of vectors in the training data set, y_i is either 1 or −1 indicating the class to which the vector x_i belongs, and b is a coefficient which determines the axis intercepts.

The geometrical margin (the separating region lies from both sides of the hyperplane) is expressed by $1/\|W\|$ (note that the maximization of $1/\|W\|$ is equivalent to the minimization of $\|W\|^2$). The slack variable S_i holds the misclassification error of the vector x_i. A slack equals 0 for a well-classified vector (not contributing to the error), and a positive value for a misclassified vector (according to the error distance).

The first part of the objective function in Eq. (1) expresses the geometrical margin maximization, where the second part describes the classification errors minimization. The slack variables range between 0 and 1 for the vectors lying within the separating area, and higher than 1 for the misclassified ones. The slack value for a well-classified vector is 0, thus not contributing to the accumulated error. The vectors with positive corresponding slacks are entitled support vectors. C is a constant which sets the relative importance of maximizing the margin versus minimizing the slack variables. A higher value of C implies a higher penalty for each misclassified vector, thus reducing the number of support vectors. Accordingly, a higher C value decreases the margin area. The constraints assure that all vectors will be classified correctly, excluding the support vectors.