Computers and Chemical Engineering 108 (2018) 152–162

Contents lists available at ScienceDirect

Computers and Chemical Engineering

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

Sensor placement for classifier-based leak localization in water distribution networks using hybrid feature selection

Adrià Soldevila[a,*], Joaquim Blesa[b], Sebastian Tornil-Sin[a,b], Rosa M. Fernandez-Canti[a], Vicenç Puig[a,b]

[a] Research Center for Supervision, Safety and Automatic Control (CS2AC), Rambla Sant Nebridi, s/n, 08022 Terrassa, Spain
[b] Institut de Robòtica i Informàtica Industrial (CSIC-UPC), Carrer Llorens Artigas, 4-6, 08028 Barcelona, Spain

ARTICLE INFO

Article history:
Received 20 May 2017
Received in revised form 1 August 2017
Accepted 3 September 2017
Available online 9 September 2017

Keywords:
Sensor placement
Leak localization
Water distribution networks
Feature selection
Genetic algorithms
Classifiers

ABSTRACT

This paper presents a sensor placement approach for classifier-based leak localization in water distribution networks. The proposed method is based on a hybrid feature selection algorithm that combines the use of a filter based on relevancy and redundancy with a wrapper based on genetic algorithms. This algorithm is applied to data generated by hydraulic simulation of the considered water distribution network and it determines the optimal location of a prespecified number of pressure sensors to be used by a leak localization method based on pressure models and classifiers proposed in previous works by the authors. The method is applied to a small-size simplified network (Hanoi) to better analyze its computational performance and to a medium-size network (Limassol) to demonstrate its applicability to larger real-size networks.

1. Introduction

Water distribution networks (WDNs) are critical infrastructures that need to be monitored to guarantee their satisfactory operation. One of the most common and critical issues to monitor and deal with are water leaks, which are account up to 30% of the total amount of extracted water (Puust et al., 2010).

Water utilities have the common practice to divide the WDN into small and non-connected areas, called District Metered Areas (DMAs), to allow a better leak monitoring and pressure control, where the inlets are monitored with flow and pressure sensors and also a few pressure sensors are placed inside. Leak localization methods rely on the use of measurements provided by a set of installed sensors. Pressure sensors are normally preferred over flow sensors because they are cheaper and easier to install and maintain.

Several leak localization methods have been proposed in the literature, such as transient analysis, parameter estimation techniques, leak sensitivity analysis, mass-balance and linear programming algorithms (Mulholland et al., 2014), statistical interval estimation (Kim et al., 2016) and artificial intelligence based methods. Artificial intelligence techniques seem to be suitable tools to use since the problem to be solved presents several types of uncertainties. For instance, in Wu and Sage (2006) genetic algorithms are proposed to solve an optimization problem for simultaneously quantifying and locating water losses. In Mashford et al. (2009), a method based on the use of Support Vector Machines (SVM) is proposed that analyzes data obtained by a set of pressure control sensors of a pipeline network to locate and compute the size of a possible leak present in a WDN. More recently, the use of k-Nearest Neighbors (k-NN), Bayesian and neuro-fuzzy classifiers for leak localization purposes has been proposed in Soldevila et al. (2016a), Soldevila et al. (2017), and Wachla et al. (2015).

Even for pressure sensors and due to budget constraints, the number of sensors that can be installed in practice is really limited. In this situation, the problem of sensor placement, i.e. the determination of the best locations inside the network to install the limited number of allowed sensors, is of utmost importance. Sensor placement in WDN was initially focus on water quality monitoring and it is still an active area of research (Rico-Ramirez et al., 2007; Chang et al., 2012; Mukherjee et al., 2017) but in the last years some sensor placement methodologies for leak localization purposes have been proposed. Examples of these sensor placement methods are presented in Sarrate et al. (2014a), where an efficient branch and bound search is used, in Casillas et al. (2013) and Cugueró-Escofet et al. (2017), where GAs are used, and in Blesa et al. (2016), where a prior clustering process is applied. In general, a given sensor placement method is designed for a particular leak localization method, since there is not a unique optimal set of sensors for a given net-

* Corresponding author.
E-mail address: adria.soldevila@upc.edu (A. Soldevila).

http://dx.doi.org/10.1016/j.compchemeng.2017.09.002
0098-1354/

work (a set of sensors can be optimal for a given leak localization method but not for a different one).

In previous works (Soldevila et al., 2016a, 2017), the authors have proposed a framework for leak localization based on computing pressure residuals, i.e. differences between measurements provided by installed sensors and estimations computed from a normal-operation model of the network, and analyzing them by a classifier. In particular, the use of the k-NN classifier is presented in Soldevila et al. (2016a), whereas the use of a Bayesian classifier is proposed in Soldevila et al. (2017). In these works, it is assumed that there exist a small number of pressure sensors that are already installed in some internal nodes of the network. In this paper, it is assumed that the number of pressure sensors to install is given and the aim is to determine their optimal locations, i.e. the ones that maximize the leak localization performances obtained when the framework proposed in Soldevila et al. (2016a) and Soldevila et al. (2017) is applied.

In this work, the problem of sensor placement is formulated as a Feature Selection (FS) problem. Feature (or variable/attribute) selection techniques (Guyon and Elisseeff, 2003) are used to identify a subset of relevant variables in a data set, regarding its use to build a model with a given purpose, for instance a classifier. Withing the framework proposed in Soldevila et al. (2016a) and Soldevila et al. (2017), the main idea is to generate, using a hydraulic simulation of the considered WDN, a complete data set containing all the potential residuals associated to the network nodes and apply a FS algorithm that determines the ones that after training the classifier will provide the best leak localization results.

There are four main categories of FS techniques recognized in the literature (Saeys et al., 2007; Bolón-Canedo et al., 2013): filter based methods, wrapper methods, embedded methods and, finally, hybrid methods, i.e. combination of filters with wrappers. The methods of the first type, filter based methods (Vergara and Estévez, 2014), directly work with the data, without interacting in any way with the model to be built. Hence, individual features or feature sets are evaluated according to some metrics that are assumed to be fast to compute. Some of the most common indicators are the relevance, i.e. the information contained in a given feature (according to the final application) (Guyon and Elisseeff, 2003; Chandrashekar and Sahin, 2014), and the redundancy, i.e. how much of the information in a given feature is repeated in others (Salmerón et al., 2016; Liu et al., 2016). Many existing filter methods combine these two indicators (Yu and Liu, 2004; Peng et al., 2005). The main advantage of this type of methods is their low computational cost, while the main drawback is that the selection does not take into account the posterior use of the data by the model. The second type of methods, wrapper methods (Chandrashekar and Sahin, 2014), build and use the model to score selected feature subsets that are generated within the framework of an heuristic search. Some methods in this category are based on the use of Genetic Algorithms (GAs) (Oreski and Oreski, 2014) and on Particle Swarm Optimization (PSO) methods (Xue et al., 2013), among others. Due to the search and to the fact that a new model has to be trained (build) for each subset, these methods are computationally demanding, but they usually provide the best results for the particular type of model used. Embedded methods are the third type of methods, and they combine the use of the model that ranks the features in a priority order to be selected. In this group, there are techniques such as Backward Feature Selection (BFS) (Guyon and Elisseeff, 2003), Random Forest (RF) (Díaz-Uriarte and De Andres, 2006) and, in general, Evolutionary Algorithms (EA) (Xue et al., 2016). Finally, the most recent approaches are the hybrid methods, which typically combine a filter that reduces the initial number of features with a wrapper that provides an additional refinement (Inbarani et al., 2014; Hu et al., 2015; Apolloni et al., 2016). The latter approach is considered in the present work due to the obtained good compromise between optimality and computation time.

Table 1
Nomenclature for physical variables.

n_n	Number of consumer nodes
$\tilde{d}_{WDN}$	Measured global demand
d_i, $\hat{d}_i$ and $\bar{d}_i$	Actual, estimated and generated demand at node i
$\mathbf{c}$, $\tilde{\mathbf{c}}$ and $\bar{\mathbf{c}}$	Actual, measured and generated boundary conditions
$\tilde{\mathbf{p}}$ and $\hat{\mathbf{p}}$	Measured and estimated inner pressures
l_i, $\hat{l}_i$ and $\bar{l}_i$	Actual, estimated and generated leaks at node i
$\mathbf{v}$ and $\bar{\mathbf{v}}$	Actual and generated noises

Table 2
Nomenclature for classifiers and feature selection.

n_c	Number of classes of each feature
n_f and $n_f^{(R)}$	Original and reduced number of features
n_s	Number of features to be selected (inner pressure sensors to be installed)
n_b	Number of fixed additional features (measured boundary conditions)
m_T and m_V	Number of instances (examples) in each class in the training and validation data sets
$\mathbf{F}$ and $\mathbf{F}^{(R)}$	Original and reduced features space
$\mathbf{T}$ and $\mathbf{T}^{(R)}$	Original and reduced training data set
$\mathbf{V}$ and $\mathbf{V}^{(R)}$	Original and reduced validation data set
$\boldsymbol{\Gamma}$	Confusion matrix
$\mathbf{D}$	Topological distance matrix
$\boldsymbol{\Phi}$ and $\boldsymbol{\Phi}^{(B)}$	Original and binarized feature distance matrix
α	Average value of the $\boldsymbol{\Phi}$ matrix except the diagonal values
σ	User defined threshold
γ	User defined threshold
$\boldsymbol{\Lambda}$	Average training matrix
p_s	Population size
e_c	Elite count parameter
tol	Fitness function tolerance
max_g	Maximum number of generations

According to the previous discussion, the contribution of this paper is the proposal of a sensor placement approach for classifier-based leak localization in water distribution networks that uses a particular hybrid feature selection algorithm, designed to reduce the computation time (for real WDNs with thousands of nodes the required computation time could be days or even weeks) while maintainig the (sub)optimality of the obtained results.

The rest of the paper is organized as follows. Section 2 presents the background, reviewing the architecture and methodology for leak localization based on pressure residuals and classifiers originally presented in Soldevila et al. (2016a) and Soldevila et al. (2017). Section 3 presents the formulation of the sensor placement problem as a feature selection problem. Section 4 details the proposed feature selection algorithm, which implements a hybrid method that uses a filter based on relevancy and redundancy/distance indicators and a wrapper based on a genetic algorithm. Section 5 presents the application of the proposed method to two networks of small and medium size: the simplified Hanoi WDN and a DMA of the Limassol WDN. Finally, Section 6 draws the main conclusions of the work.

1.1. Nomenclature

The names for the main variables and parameters used through the paper are summarized in Tables 1 and 2.

2. Background: leak localization based on pressure residuals and classifier

2.1. Architecture and operation

In a previous work (Ferrandez-Gamot et al., 2015), the authors proposed an on-line leak localization method that relies on the