2014

Spectral Clustering And Support Vector Classification For Localizing Leakages In Water Distribution Networks – The ICeWater Project Approach

Antonio Candelieri

Dante Conti

Davide Soldi

Francesco Archetti

11th International Conference on Hydroinformatics
HIC 2014, New York City, USA

SPECTRAL CLUSTERING AND SUPPORT VECTOR CLASSIFICATION FOR LOCALIZING LEAKAGES IN WATER DISTRIBUTION NETWORKS – THE ICEWATER PROJECT APPROACH

CANDELIERI A. (1,2), CONTI D. (2), SOLDI D. (1,2), ARCHETTI F. (1,2)
(1): Department of Computer Science, Systems and Communication, University of Milano-Bicocca, Italy
(2): Consorzio Milano Ricerche, via Roberto Cozzi 53, Milano, Italy

This paper presents an analytical framework aiming at improving the leakage management process in Water Distribution Network (WDN). The approach uses: 1) hydraulic simulation software (EPANET) to run several "leakage scenarios", by varying leak location (pipe) and severity and to store variations in pressure, at nodes, and flow, on pipes; 2) Spectral Clustering, which is applied on a graph, generated from the simulation data, having scenarios as nodes and edges weighted by the similarity between each pair of scenarios, in terms of pressure and flow variations; the goal is to group together leaks implying similar variations; 3) Support Vector Machine classification learning to discover a relation linking variations in pressure and flow to a limited set of probably leaky pipes. The approach also proposes a strategy to support cost-effective placement of flow and pressure meters, through the identification of the best trade-off between reliability in localization and deployment costs. As a result, the overall approach has been validated on the Italian pilot site of the European project ICeWater, offering a very high reliability in localizing leak: about 98%, both on training and test data.

INTRODUCTION

Nowadays urban Water Distribution Networks (WDNs) suffer leakage, mainly due to the age of the infrastructures, implying failures, large amounts of Non Revenue Water (NRW) and high costs for energy (i.e., pumping) and rehabilitation, while budgetary constraints are becoming more strong. The International Water Association (IWA) highlighted the relevance to improve the leakage management process [1] which is usually divided in three consecutive steps [2]: assessment, detection and physical localization.
Several studies proposed to improve localization through the analysis of data collected by computer-based systems usually adopted in WDNs, such as Supervisory Control And Data Acquisition (SCADA), Automatic Metering Readers (AMR), GIS and hydraulic simulation software. Many approaches use machine learning, statistics, probabilistic modeling have been investigated [3-6], with the common idea that actual modifications in flow and pressure within the WDN are linked to a set of leaky pipes: hydraulic simulation software may be therefore used to simulate a wide set of leaks, store variations and then discover the inverse relation

between variations and leaky pipes [7-9]. While most of the proposed approaches try to localize a leak on pipes, in [10] a combination between hydraulic simulation and classification learning has been developed to identify leaks on junctions.
This paper presents an analytical framework that uses: 1) extensive simulation of leaks for data generation, 2) network-based Spectral Clustering to group together leaks implying similar variations in pressure and flow, 3) classification learning (i.e., Support Vector Machine, SVM) to discover the relation linking variations in pressure and flow to a limited set of probably leaky pipes (i.e., a cluster of those provided by Spectral Clustering). The approach also proposes a strategy to support cost-effective placement of flow and pressure meters, identifying the best trade-off between reliability in localization and deployment costs.
All the results are related to a real test case, a Pressure Management Zone (PMZ) of the WDN in Milan, Italy, one of the two pilots of the FP7-ICT project ICeWater co-funded by the European Commission.

MATERIALS AND METHODS

Description of the pilot and the data generation process

In the following Figure 1 the ICeWater pilot "Abbiategrasso", in Milan, Italy, is depicted. The Figure has been obtained by using EPANET, a hydraulic simulation software widely used for modeling WDNs and downloadable for free from the Environmental Protection Agency web site (http://www.epa.gov/nrmrl/wswrd/dw/epanet.html). EPANET permits to perform what-if simulation and it can be also integrated with optimization models for supporting decisions both at operational, planning and strategic level.

Figure 1. The PMZ "Abbiategrasso", the ICeWater's pilot in Milan, Italy.

Abbiategrasso is a PMZ consisting of 1212 junctions (612 are consumption points) and 1385 pipes; the overall pipe infrastructure is long about 116905m. Pipes length ranges from 0.25m to 844.92m (average 84.41m), pipes diameter ranges from 50mm to 900mm (average 244.92mm).
In this study EPANET has been used to simulate a wide set of "leakage scenarios" where a scenarios is obtained by placing, in turn, a leak on each pipe of the network model and varying its severity in a given range. At the end of each run, the corresponding scenario is represented by the variations, in pressure (at junctions) and flow (at pipes), computed with respect to the faultless network. These simulation results are stored in a dataset together with the information related to the leaky pipe and the leak severity. More details about the pressure-dependent leak modeling have been firstly described in [7].

Clustering Leakage Scenarios and Quality Measures

The main step of the proposed approach consists in grouping together scenarios (rows of the dataset) that are similar in terms of variations in pressure and flow induced by the leak, while information on leaky pipe and leak severity is ignored. This step has been named "clustering leakage scenarios". As many clustering algorithms are available, a measure has to be used to evaluate the quality of the solution with respect to the final goal. In this case, the main aim is to obtain clusters of scenarios that are related to restricted sets of pipes. Respect to this, although several (internal) indexes for evaluating the validity of clustering algorithms are available, *ad-hoc* indexes had to be defined by authors.

The first index measures how much the identified clusters are associated to restricted sets of WDN's pipes. Namely, the *Localization Index* of the cluster *k* is obtained by retrieving the information about the leaky pipe of each scenario in the cluster *k* and is then computed as:

$$LI_k = \frac{|pipes| - |pipes_k|}{|pipes| - 1} \tag{1}$$

where | *pipes* | is the overall number of pipes of the WDN and | $pipes_k$ | is the number of leaky pipes of the scenarios into cluster *k*.

The maximum value of LI_k is $LI_k = 1$ that is obtained when the cluster *k* contains scenarios with leak on only one pipe (i.e., | $pipes_k$ | = 1); the minimum value is $LI_k = 0$ that is obtained if the scenarios in the cluster *k* are associated to all the pipes of the WDN (i.e., | $pipes_k$ | = | *pipes* |).

The overall LI of a clustering procedure is given by the average of the LI_k weighted by the number of distinct pipes in each cluster.

The second proposed index measures how much obtained clusters contain scenarios related to same (leaky) pipe, even if with different leak severity. Namely, *Quality of Localization* of the cluster *k* is defined as:

$$QL_k = \frac{\sum_{p \in pipes_k} \frac{n_p^k}{|S|}}{|pipes_k|} \tag{2}$$

where *S* is the set of different severities used (and |*S*| is the overall number of severity values used), *p* is a distinct pipe within cluster *k*, n_p^k is the number of scenarios in cluster *k* and associated to the pipe *p*.

The maximum value of QL_k is $QL_k = 1$ that is obtained when the cluster *k* contains all the scenarios related to the each pipe in the $pipes_k$ set.

The overall Quality of Localization for a clustering procedure is given by the average of QL_k.

Finally, a global index LI* is defined, combining *LI* and *QL*:

$$LI^* = LI \times QL \tag{3}$$

The maximum of *LI** is *LI** = 1 and it is obtained only when *LI* and *QL* are equal to 1.

In this study, network- and not-network- ("traditional") based clustering techniques have been investigated. In particular, *graph clustering* [11] has been proposed to group nodes of a graph into sub-graphs (clusters) maximizing the sum of the weights on the edges within each cluster (intra-cluster similarity) while minimizing the sum of the weights on the edges connecting nodes in different clusters (inter-cluster similarity).In order to apply graph clustering, in this study leakage scenarios have been represented as nodes of a graph and edges have been

weighted by the cosine-similarity [12] between two scenarios (i.e., two vectors of variations in pressure and flow). The proposed approach is based on Spectral Clustering [13, 14], whose core consists in the eigen-decomposition of a n x n matrix, with n the number of nodes of the graph. Different alternative definitions have been proposed and studied through graph theory [15]; the usually adopted *Normalized Laplacian* is:

$$L_{norm} = I - D^{-1/2} A D^{-1/2} \tag{4}$$

where A is the affinity matrix, whose entries are the similarities computed for each pair of nodes of the undirected graph, and D is the degree matrix, having the degree of nodes on the diagonal and 0 on the other entries.

The process to obtain K cluster is usually performed by representing data in the space spanned by a restricted set of relevant eigenvectors of the (Normalized) Laplacian matrix [16, 14]. When a similarity measure is used to built the Affinity matrix, the relevant eigenvectors to are the first l smallest. To select the most appropriate l, eigenvalues are sorted in ascending order and they are selected as relevant until eigengap (i.e., difference between two successive eigenvalues) lower than a given value. One of the widely used implementations of Spectral Clustering is that proposed in [17], consisting in selecting the l smallest non-zero eigenvalues and performing a traditional k-means clustering of the nodes in the eigen-space. Anyway, any other traditional clustering algorithm may be applied in this space, as reported in the section on results.

Support Vector Machines to identify a restricted set of leaky pipes

After clustering the leakage scenarios, the next step consists in discovering a reliable relation between the variations in pressure and flow, due to a leak, and the correspondent cluster, which permits to retrieve the set of correspondent leaky pipes. Spectral Clustering procedure implicitly uses a non-linear mapping from the space related the variations, in pressure and flow, to the space spanned by the most relevant eigen-vectors of the Laplacian, that can be approximated by supervised machine learning approaches, such as Support Vector Machines (SVM), as proposed in this study. More in detail, input of the classification learning strategies are the leakage scenarios, that are vectors of variations in pressure and flow at the sensors, while output (class attribute) is the cluster provided by the Spectral Clustering procedure.

Cost-effective sensors placement

Since variations in pressure and flow at the monitoring points are input both of Spectral Clustering and SVM classification, number and position of sensors affects overall performance of the approach proposed. Ideally, the greater the number of deployed sensors the higher is the quality of clustering and accuracy of the SVM classifier. However, deployment implies high costs for equipment and installation as well as useless redundancy of information.

Optimal sensors placement has been recently addressed for both leakage detection/location [18, 19] and water quality issues [20, 21]. In this study, a solution for cost-effective sensor placement is proposed, aimed at identifying the best trade-off between high reliability in leakage localization (effectiveness) and costs for sensors. The solution uses, again, clustering on the dataset of leakage scenarios; in this case clustering (Partitioning Around Medoids, PAM) is applied on the columns of the dataset, that are variations in pressure and flow at each junction and each pipe, respectively. Clustering is performed separately for junctions and pipes sets,

with the aim to group together, separately, those that are similar over the simulated leakage scenarios. Only the medoids of the clusters are selected as the most relevant monitoring points (that are a pressure meter in the case of junction medoid and a flow meter in the case of pipe medoid).

EXPERIMENTAL RESULTS

In this section, the results are presented. The overall number of leakage scenarios that have been generated is 29800 (divided in 50% training and 50% test set), obtained by placing a leak, in turn, on each pipe, and varying its severity among 10 different values.

Results on cost-effective sensors placement

In the following Figure 2 the results of all the sensors placements considered are depicted: 7, 10, 13 or 16 pressure meters and 3, 4, 5 or 6 flow meters. *LI** has been considered as global index of localization reliability (effectiveness), while costs for sensors have been set 1 for a pressure meter and 10 for a flow meter.
Planning a sensor placement of 10 pressure and 5 flow meters appears to be, globally (LI*), the best choice in terms of trade-off between leakage localization and deployment costs.

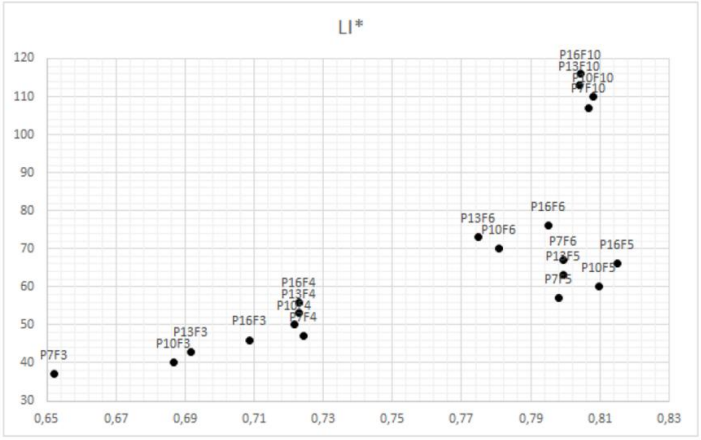

Figure 2. Costs for sensors (y-axis) versus LI* (x-axis).

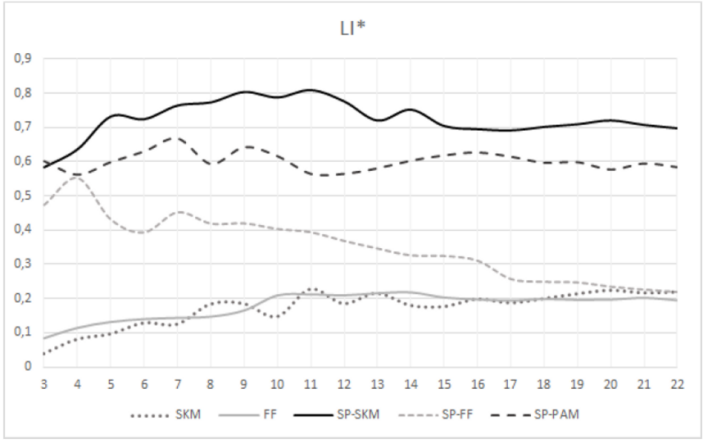

Figure 3. LI* depending on number of clusters (K), traditional versus network-based clustering algorithms: simple K-means (SKM), Farthest First (FF), Spectral Clustering (SP).

Results on Spectral Clustering

This section reports a comparison among three different implementations of Spectral Clustering (i.e., internally using: a) simple K-means, b) Farthest-First and c) PAM) and two "traditional", not-network-based clustering algorithms (i.e., simple K-means and Fartest-First).
The previous Figure 3 shows the trend of LI* with respect to the number of desired clusters for each algorithm. Taking into account the definition of LI, it is quite easy to understand that increasing K improves the Localization Index (LI) while reduces the QL.
The best configuration selected in this paper is the algorithm S-SKM with K=11 (and 3 eigen-vectors). Spectral Clustering performances are clearly higher than those offered by clustering algorithms which are not-network-based. Finally, in order to give an idea about the width of regions where physically check for a leak, the following Figure 4 shows the best and the worst clusters in terms of LI*.

Figure 4. Set of probably leaky pipes in one of the most (left) and less (right) localizing clusters (result from the best Spectral Clustering).

Results on SVM-based leaky pipes identification

The implementation of C-SVM provided in the open-source Java-based suite WEKA (Waikato Environment for Knowledge Analysis, http://www.cs.waikato.ac.nz/ml/weka/) has been adopted. In particular, the Radial Basis Function (RBF) kernel has been adopted for non-linear mapping. Both *C* and the internal parameter γ of the kernel have been varied until accuracy, on 10 fold-cross validation, is no more improved. Accuracy is the percentage of vectors correctly associated to the cluster provided by the Spectral Clustering process; 10 fold-cross validation technique uses the entire dataset to train a model and test it, giving an estimation of the reliability in predicting the class label (i.e., cluster associated to new vectors of hydraulic variations, in this case). The best SVM configuration resulted setting $C = 1$ and $\gamma = 1$. The learned SVM classifier has been then validate on an independent test set, related to leakage scenarios obtained on values of severity different from those already adopted (i.e. new leaks). Neither Spectral Clustering or SVM training are performed on this test set; the vectors of pressure and flow variations, associated to a leak, are given as input to the learned SVM classifier, which provides an estimation of the cluster probably assigned by Spectral Clustering. If the leaky pipe associated to the specific vector of variations is in the set of distinct pipes associated to the predicted scenarios cluster a successful localization is counted for.
The following Figure 5 summarizes the performances related to the number of successful localizations, both on training (97.99%, average) and independent test (98.02%, average).

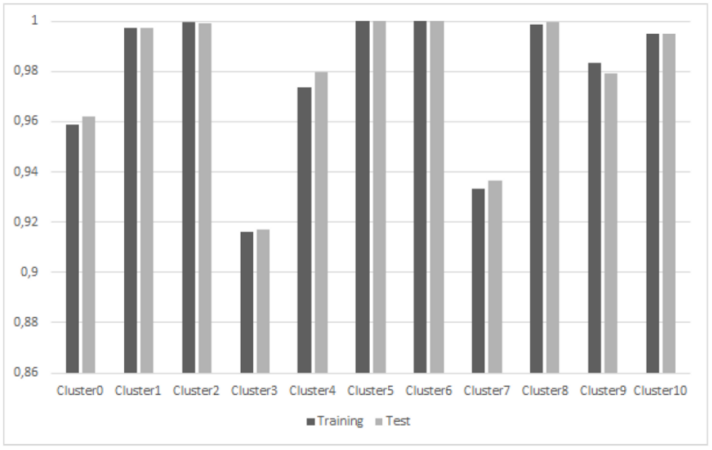

Figure 5. Successful leakage localization, both on training and test set and for each cluster.

CONCLUSIONS

The approach presented in this paper aims at improving leakage localization in urban WDN through simulation of several leakage scenarios, Spectral Clustering and Support Vector Machine classification. A reliable relationship (reliability about 98%) between variations, in pressure and flow, and leak location has been identified and can be used to reduce time and costs for investigations and rehabilitation: when a leak is detected (e.g., with traditional methods, such as Minimum Night Flow analysis [3]) actual pressure and flow measurements are given as input to the SVM which provides the set of probably leaky pipes (cluster of the Spectral Clustering) associated to that variation. The framework supports also cost-effective sensor placement. The overall approach has been validated on a real case study, the Abbiategrasso PMZ, in Milan, Italy, one of the two pilots of the European project ICeWater.

ACKNOWLEDGEMENTS

This work has been partially supported by the European Union ICeWater project – FP7-ICT 317624 (www.icewater-project.eu).

REFERENCES

[1] *Alegre H., Baptista J.M., Cabrera E., Cubillo F., Duarte P., Hirner W., Merkel W. and Parena, R., "Performance Indicators for Water Supply Services", Second Edition, IWA Publishing (2006).*

[2] *Puust R., Kapelan Z., Savic D. A. and Koppel, T., "A review of methods for leakage management in pipe networks". Urban Water Journal, Vol. 7, No. 1 (2010), pp 25-45.*

[3] *Behzadian K., Kapelan Z., Savic D. A. and Ardeshir A., "Stochastic sampling design using multi objective genetic algorithm and adaptive neural networks", Environmental Modeling and Software, Vol. 24, (2009), pp 530–541.*

[4] *Xia L., and Guo-jin L., "Leak detection of municipal water supply network based on the cluster-analysis and fuzzy pattern recognition", International Conference on E-Product E-Service and E-Entertainment (ICEEE), Vol. 1 No. 5, (2010), pp 7-9.*

[5] *Nasir A., Soong B. H., Ramachandran S., "Framework of WSN based human centric cyber physical in-pipe water monitoring system", Proc. 11th International Conference on Control, Automation, Robotics and Vision, (2010), pp 1257-1261.*

[6] *Lijuan W., Hongwei Z. and Hui J., "A Leak Detection Method Based on EPANET and Genetic Algorithm in Water Distribution Systems", Software Engineering and Knowledge Engineering: Theory and Practice – Advances in Intelligent and Soft Computing, Vol. 14, (2012), pp 459-465.*

[7] *Candelieri A. and Messina E., "Sectorization and analytical leaks localizations in the H2OLeak project: Clustering-based services for supporting water distribution networks management", Environmental Engineering and Management Journal, Vol. 11 No. 5, (2012), pp 953-962.*

[8] *Candelieri A., Conti D. and Archetti F., "A graph based analysis of leak localization in urban water networks", Proc. 12th International Conference on Computing and Control for the Water Industry, CCWI201, (2013).*

[9] *Candelieri A., Archetti F. and Messina E., "Improving leakage management in urban water distribution networks through data analytics and hydraulic simulation". WIT Transactions on Ecology and the Environment, Vol. 171, (2013b), pp 107-117.*

[10] *Mashford J., De Silva D., Burn S. and Marney D., "Leak Detection in simulated water pipe networks using SVM", Applied Artificial Intelligence: An International Journal, Vol. 26 No. 5, (2012), pp 429-444.*

[11] *Schaeffer S.E., "Graph Clustering (survey)", Computer Science Review, (2007), pp 27-64.*

[12] *Tan P.N., Steinbach M. and Kumar V., "Introduction to Data Mining", Addison-Wesley, (2005).*

[13] *Jaakkola T., "Course materials for 6.867 Machine Learning", Fall 2006. MIT OpenCourseWare (http://ocw.mit.edu/), Massachusetts Institute of Technology, (2006).*

[14] *Luxburg U., "A Tutorial on Spectral Clustering. Statistics and Computing", Vol. 17 No. 4, (2007), pp 1-32.*

[15] *Chung F., "Spectral graph theory", Washington: Conference Board of the Mathematical Sciences, (1997).*

[16] *Ng A.Y., Jordan M. and Weiss Y., "On Spectral Clustering: Analysis and an algorithm", Advances in Neural Information Processing Systems, Vol. 14, (2001), pp 849-856.*

[17] *Shi, J., Malik, J., 2000. Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(8), 888-905.*

[18] *Christodoulou, S.E., Gagatsis A., Xanthos S., Kranioti S., Agathokleous A. and Fragiadakis M., "Entropy-Based Sensor Placement Optimization for Waterloss Detection in Water Distribution Networks", Water Resour Manage, Vol. 27, (2013), pp 4443–4468.*

[19] *Casillas M.V., Puig V., Garza-Castanon L.E. and Rosich A., "Optimal Sensor Placement for Leak Location in Water Distribution Networks Using Genetic Algorithms", Sensors Vol. 13, (2013), pp 14984-15005.*

[20] *Klise K., Phillips C. and Janke R., "Two-Tiered Sensor Placement for Large Water Distribution Network Models", Journal of Infrastructure Systems, Vol. 19 No. 4, (2013), pp 465–473.*

[21] *Chang N.-B., Pongsanone N.P., Ernest A., "A rule-based decision support system for sensor deployment in small drinking water networks", Journal of Cleaner Production, Vol. 29–30, (2012), pp 28-37.*