Clustering-based Burst Detection Using Multiple Pressure Sensors in District Metering Areas

Yipeng Wu[1], Shuming Liu[2]
[1,2] School of Environment, Tsinghua University, 100084, Beijing, China
[2]*shumingliu@tsinghua.edu.cn*

ABSTRACT

Bursts in water distribution systems (WDS) can cause water loss, service interruptions and other negative effects. However, it is challengeable for worldwide water utilities to timely be aware of bursts. This paper presents a novel burst detection approach using data from multiple pressure sensors in district metering areas (DMA). Differing from most data-driven methods that employ prediction models, this method utilizes a clustering algorithm to detect burst-induced data. Owing to the use of cosine distance in clustering analysis, temporal varying correlation between data from diffierent sensors is exploited, making the method only requires one day's worth of data to implement. When applied to a DMA with three pressure sensors, the method successfully detected some real and simulated bursts over a period of two months.

Keywords: Burst Detection; Cosine Distance; Clustering.

1 INTRODUCTION

Worldwide water utilities have been troubled by leakage in water distribution systems (WDSs) for decades. Significant loss of treated water and huge cost of repairs caused by leakage have exerted tremendous financial pressure on utilities. Pipe bursts are a form of leakage characterized by short duration but typically high flow. Other than the waste of water, the negative effects of bursts, such as customer service interruptions and the intrusion of contaminants through broken pipes, are non-negligible [1, 2]. In order to facilitate water utilities' quick reaction to pipe bursts, effective and efficient methods are required to be aware of bursts in a timely manner.

Owing to use of the supervisory control and data acquisition (SCADA) system, near real-time hydraulic data are collected so that network behaviours including bursts can be identified with data-driven detection approaches [3]. Some of these methods have been applied in real-life district metering areas (DMA) and showed respectable detection performance [4]. However, most of existing data driven approaches are based on a prediction-classification two-stage framework [1]. The prediction stage estimates ideal data under normal network conditions using prediction models, such as artificial neural networks (ANN), Kalman filter and polynomial models [4-6]. When it comes to the classification stage, difference between predicted and observed hydraulic values is evaluated. If the difference is large enough, an alarm will be triggered to indicate a burst or an event. Consequently, the accuracy of classification is determined by the prediction stage. In order to fit variations of hydraulic data (e.g., diurnal pattern of flow readings), vast historical data (extending many weeks and sometimes years) are needed to train above-mentioned models. Noticeably, because of the existence of unusable historical data (e.g., event-induced data, missing data and replicated data), data pre-processing is indispensable before the training of prediction models. The accuracy of prediction is prone to being affected by the quality of this process. Furthermore, prediction-classification methods cannot immediately be applied to newly built or reformed

networks due to the shortage of historical data. To overcome the limitations, it is necessary to develop new methods that require much less historical data (e.g., a day's worth of data).

This paper presents a burst detection method based on a clustering algorithm. The approach uses data from multiple pressure sensors to identify whether a burst occurs in a DMA near real time. Compared with most current data-driven approaches, it clusters actual measurements using the dissimilarity between pressure measurements rather than analysing discrepancies between monitored and predicted values. Most importantly, this method only requires a day of time series data to implement because it fully utilizes temporal varying correlation between pressure data from different sensors. The methodology is discribed in section 2. In section 3, some real events and simulated bursts in a DMA located in southern China are used to evaluate the performance of this method.

2 METHODOLOGY

2.1 Data Transformation

Figure 1 shows the typical diurnal patterns of three pressure sensors in a DMA. A burst occurs between 8:40 and 9:00. It is clear that the data from different sensors fluctuate greatly, though, they always vary at the same time with similar changes in amplitude. This feature between data from multiple sensors is defined as temporal varying correlation.

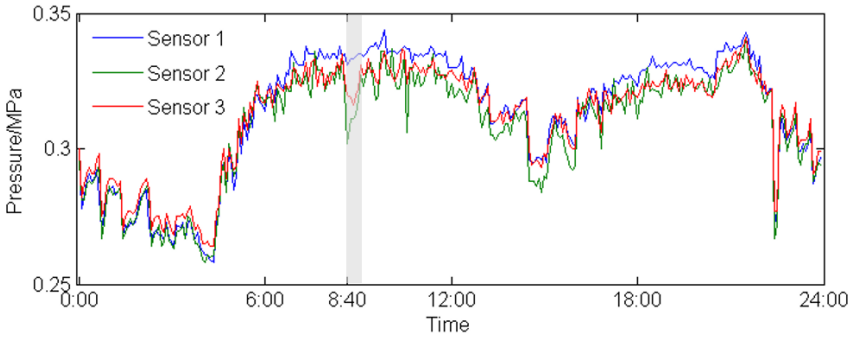

Figure 1. One day's worth of pressure data from three sensors in a DMA. The data within the grey area represent a burst event.

Pressure data drop significantly for two of the sensors, whereas the other sensor's data only change slightly during the burst event. In other words, the temporal varying correlation disappears due to the burst. In a DMA, the value ranges of pressure sensors may differ from each other. To reinforce the temporal varying correlation and make the burst-induced data more distinguishable, the time series from each pressure sensor is divided by its median value for normalization (shown in Figure 2). Then the normalized data are transformed into a matrix. Each row in the matrix is a vector that includes 3 normalized measurements from different sensors at the same time step.

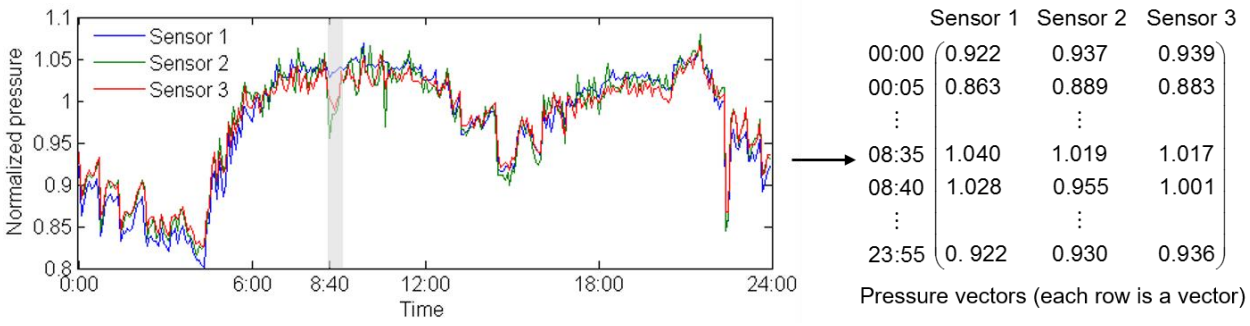

Figure 2. Normalized pressure data and pressure matrix.

2.2 Cosine Distance and Clustering Algorithm

To autumatically identify burst-induced vectors (or some abnormal vectors caused by other events), this paper employs a clustering algorithm developed by Rodriguez and Laio [7]. A proper dissimilarity measure is an important issue in clustering analysis and the selection of it should consider features of the data to be analysed. As shown in Figure 3A, most of 288 vectors from the pressure matrix are approximately along the same direction in 3-dimensional space. This is essentially the concrete embodiment of temporal varying correlation between data from multiple sensors in space. The vectors representing a burst (red points) distinguish themselves because they deviate from this direction. Considering that, cosine distance, based on the cosine of the angle between any two vectors, is used to measure the dissimilarity between each set of two vectors. When the angle between two vectors is small (i.e., the vectors have approximately the same direction), they are similar to each other even when all elements in the vectors have different magnitudes (e.g., $\boldsymbol{x}$ and $2\boldsymbol{x}$). By contrast, the two vectors are dissimilar if the angle between them is large. Consequently, normal vectors are similar to each other and are dissimilar to burst-induced vectors. At this point, the clustering algorithm can be used to differentiate vectors with high dissimilarity.

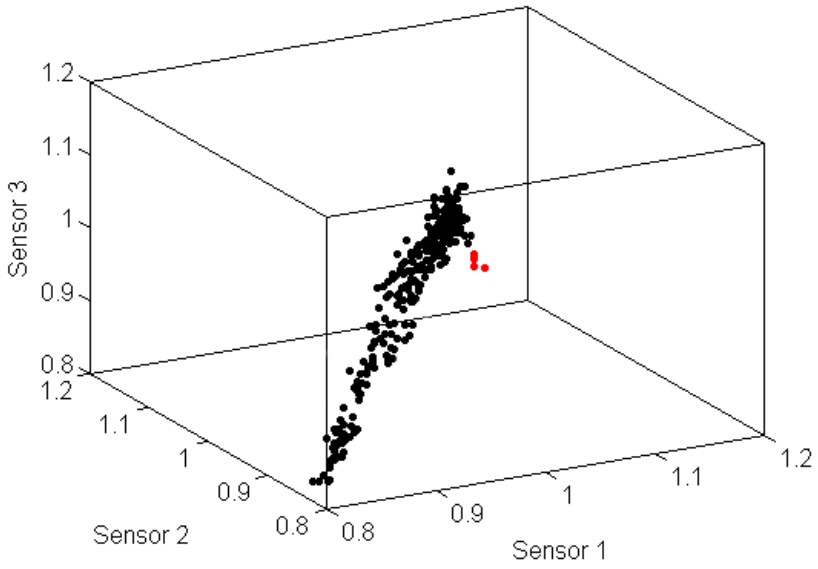

Figure 3. Location distribution of pressure vectors in 3-dimensional space. All vectors are from the matrix in Figure 2 and the red ones are burst-induced vectors.

In clustering analysis, two quantities need to be computed for each vector: the vector's local density ρ and its distance σ from vectors of higher density. The vectors that may represent bursts, with high dissimilarity to others, are deemed as outliers. Because the purpose is outlier detection, this method is not concerned with the number of clusters. Wu et al. [8] have described the calculation procedure of the two quantities in detail and defined outliers as the vectors with lowest ρ and significantly large σ. To quantify how large σ should be, a significance factor α is introduced. The vectors with the lowest local density are ranked in descending order according to the σ of each vector. If the value of α is 0.1, the top 10% of those ordered vectors are regarded as outliers. It should be noted that both ρ and σ are computed based on cosine distance between vectors in this paper. Assuming that a matrix only contains normal vectors (defined as reference matrix in the paper), the clustering algorithm can identify a newly collected burst-induced vector by evaluating the dissimilarity between it and other normal vectors.

3 CASE STUDY

The case study is located in a DMA in southern China, with an average water demand of 4500 m^3. The number of consumers is 6950 and most of them are urban residents. The DMA, supplied via a single main, is monitored by a flow sensor at the inlet. The layout of three pressure sensors used in this study is shown in Figure 4. The sampling interval of these sensors is 5 minutes. To evaluate the detection performance of the clustering-based method, data from 1 June to 31 July 2016 were collected. Missing values were replaced with zeros to ensure data continuity.

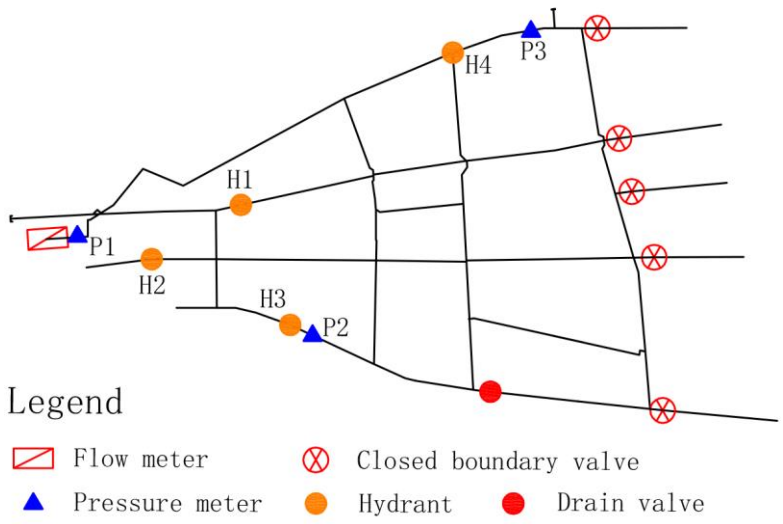

Figure 4. Pipe network of the case study.

Data for 1 June, a day on which no event occurred, are used to form a reference matrix. Pressure data from 2 June to 31 July form 17,280 3-dimensional vectors (there are 60 days and each day has 288 points in time). Every vector is to be detected in sequence by evaluating its disimilarity to the reference matrix. That is to say, every detection carries out the clustering analysis for the reference matrix and a new vector (289 vectors in total) to identify whether the new vector is an outlier. As discussed in section 2.1, all data are normalized before clustering. There are 2 incidences of hydrant damage, 2 incidences of pipe flushing and 3 burst simulations over this period and all of them can be regarded as bursts. Table 1 presents the detailed description of these events. According to the duration of the events (confirmed by the records from the water utility and flow monitoring data),

50 vectors should be detected as outliers. This is the basis to calculate true positive rate (TPR) and false positive rate (FPR).

Table 1. Detailed description of some events and corresponding detection result (α =0.5).

Type	Location	Duration	% of average daily flow	Detected outliers and corresponding time
Hydrant damage	H4	14/7/2016 11:40-12:05	70%	3(11:40-11:50)
	H2	24/7/2016 8:40-9:00	40%	5(8:40-9:00)
Simulated burst	H1	22/7/2016 14:40-14:50	120%	3(14:40-14:50)
	H2	22/7/2016 15:00-15:13	100%	2(15:00-15:05)
	H3	22/7/2016 15:24-15:30	40%	1(15:25)
Pipe flushing	Drain valve	21/6/2016 13:10-13:45	90%	6(3:10-13:35)
	Drain valve	22/6/2016 14:15-16:05	90%	21(14:15-15:55)

Figure 5 depicts the change of TPR and FPR when the significance factor α varies on the interval [0.1,1]. Differing from other detection methods with a binary result (i.e., outlier or non-outlier), the FPR of this clustering-based method cannot reach 100% and is far less than 100% no matter how large the value of α is. The reason is that the method only identifies outliers among vectors with the lowest ρ rather than among all vectors. Therefore, Figure 5 is not a typical ROC curve. The method becomes more sensitive to bursts with the use of larger α. When α is larger than 0.3, all 7 events can be identified successfully. Table 1 shows the number of detected outliers and the corresponding time at which detected outliers occur (α =0.5). Note that the TPR stays unchanged (below 100%) although the value of α increases from 0.6 to 1. As shown in the grey area of Figure 1, pressure tends to get back to normal level at the end of a burst. The corresponding vectors are similar to normal vectors so they cannot be distinguished (as shown in the detection result in Table 1).

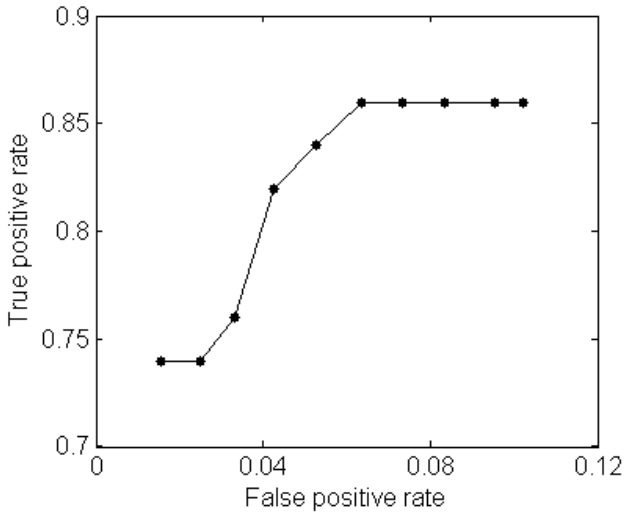

Figure 5. Change of true positive rate and false positive rate with the increase of significance facator.

FPR, which should be kept small, also becomes larger when α increases. A large FPR will cause more false alarms and make the detection method unreliable. Consequently, the value of α should be relatively small and a proper alarm rule needs to be set. When α equals 0.5, 951 vectors were identified as outliers and 86 of them were caused by missing data. The remaining 865 outliers were

classified into 3 conditions: a) detected with a single outlier; b) detected with 2 continuous outliers; c) detected with more than 2 continuous outliers. A total of 599, 126, and 140 outliers belong to conditions a, b, and c respectively. Considering the less stationary nature of pressure data [9], the large number of condition a and condition b is not unexpected. To reduce false alarms, only outliers that belong to condition c will trigger alarms. In other words, only 3 or more continuously detected outliers are regarded as bursts. Therefore, 30 alarms were raised by the 140 outliers of condition c and 5 of them were triggered by events listed in Table 1. The other 25 alarms might be caused by unexpected demand of large consumers or unrecorded events. Under this alarm rule, the method cannot raise alarms for the two burst simulations with short duration even though some outliers were detected. Furthermore, the decrease in the number of alarms is at expense of the increase of detection time.

4 CONCLUSIONS

The method proposed in this paper successfully utilised data from multiple pressure sensors to detect real and simulated bursts in a DMA. The conclusions of this work and suggestions for future work are listed below:

1. Temporal varying correlation exists between the data from multiple pressure sensors in a single DMA although the variation range of data is wide with the change in time. Cosine distance is a sutitable dissimilarity measure for this kind of time series data. It makes burst-induced data distinguishable from a normal dataset (i.e., reference matrix) which can be formed by only one day's worth of pressure data.
2. Facilitated by a proper dissmilarity measure, clustering algorithm is a useful tool to detect bursts. The clustering-based method is sensitive to large bursts (larger than 40% of average daily DMA flow) with a relatively small significance factor (e.g., 0.4-0.5). It still needs to be tested whether this method can identify relatively small bursts.
3. A pressure sensor is more sensitive to bursts that occur near it. Further work could focus on mining location information after a burst is detected using this method.

References

[1] Y. Wu and S. Liu, "A review of data-driven approaches for burst detection in water distribution systems," *Urban Water Journal*, pp. 1-12, 2017.

[2] S. Fox, W. Shepherd, R. Collins, and J. Boxall, "Experimental quantification of contaminant ingress into a buried leaking pipe during transient events," (in English), *Journal of Hydraulic Engineering*, Article vol. 142, no. 1, Jan 2016.

[3] M. Romano, Z. Kapelan, and D. A. Savic, "Automated detection of pipe bursts and other events in water distribution systems," (in English), *Journal of Water Resources Planning and Management*, Article vol. 140, no. 4, pp. 457-467, Apr 2014.

[4] S. R. Mounce, J. B. Boxall, and J. Machell, "Development and verification of an online artificial intelligence system for detection of bursts and other abnormal flows," (in English), *Journal of Water Resources Planning and Management-Asce*, Article vol. 136, no. 3, pp. 309-318, May-Jun 2010.

[5] G. L. Ye and R. A. Fenner, "Kalman Filtering of Hydraulic Measurements for Burst Detection in Water Distribution Systems," (in English), *Journal of Pipeline Systems Engineering and Practice*, Article vol. 2, no. 1, pp. 14-22, Feb 2011.

[6] G. L. Ye and R. A. Fenner, "Weighted least squares with expectation-maximization algorithm for burst detection in U.K. water distribution systems," *Journal of Water Resources Planning and Management,* vol. 140, no. 4, pp. 417-424, 2014.

[7] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," (in English), *Science,* Article vol. 344, no. 6191, pp. 1492-1496, Jun 2014.

[8] Y. Wu, S. Liu, X. Wu, Y. Liu, and Y. Guan, "Burst detection in district metering areas using a data driven clustering algorithm," *Water Research,* vol. 100, pp. 28-37, 9/1/ 2016.

[9] S. R. Mounce, R. B. Mounce, and J. B. Boxall, "Novelty detection for time series data analysis in water distribution systems using support vector machines," *Journal of Hydroinformatics,* vol. 13, no. 4, pp. 672-686, 2011.