Journal of Process Control 82 (2019) 13–21

Contents lists available at ScienceDirect

Journal of Process Control

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

Optimal sensor placement for leak localization in water distribution networks based on a novel semi-supervised strategy

Juan Li*, Cong Wang, Zhihong Qian, Changgang Lu

Jilin University, Nanhu Road, No. 5372, Changchun, 130012, China

ARTICLE INFO

Article history:
Received 10 December 2018
Received in revised form 15 July 2019
Accepted 6 August 2019
Available online 23 August 2019

Keywords:
Optimal sensor placement
Leak localization
Water distribution networks
Semi-supervised feature selection
Fuzzy *c*-means
Semi-JMI

ABSTRACT

Optimal sensor placement (OSP) techniques are important to monitor water distribution networks (WDNs). However, these methods are only feasible under the premise of complete leak location information. In fact, partial leak locations may become lost or may not be recorded. Therefore, in this case, the previous OSP methods are not applicable. To overcome this problem, this paper proposes a new OSP strategy considering that some leak locations are unknown, meaning that they are under semi-supervised conditions. The proposed semi-supervised strategy and the original semi-JMI algorithm are applied to two benchmark networks. The results indicate that the monitoring nodes selected by the semi-JMI criterion will lead to the occurrence of monitoring blind zones, resulting in the degradation of the leak localization performance. The addition of the fuzzy *c*-means (FCM) clustering method will compensate for this problem and improve the accuracy and the stability of leak localization.

1. Introduction

The water distribution network (WDN) is a civil infrastructure system that guarantees the transportation, distribution and provision of potable water in urban areas [1,2]. Due to aging and poor maintenance, WDNs lack monitoring and security. Therefore, water loss in WDNs may account for 30% of the total amount of extracted water [3]. In addition, water leakage may have serious economic and social impacts. Hence, leakage management requires novel and efficient methods to address these challenges.

In recent years, the optimal sensor placement (OSP) problem has been an active field of research, making network repair easier and faster. Most of these studies focus on contamination monitoring [4]. In addition, due to the cost constraints, several leak localization approaches have been proposed for installing a limited number of sensors in the WDN [5]. For instance, a method based on hydraulic model simulations has been proposed to identify the OSP [6]. In [7], a method combining the clustering technique with a branch and bound search based on a structural model has been utilized to deal with the OSP problem. In [8], a robust sensor placement methodology was proposed. This methodology was formulated into a multi-objective optimization strategy. In [9], a Bayesian classifier is applied to analyze the calculated pressure changes, namely, the differences between the measurements provided by the installed sensors and estimations computed from a normal-operation model of the network. In [10], a hybrid feature selection method, which combines a filter with a wrapper on the basis of genetic algorithms, is presented.

The OSP methods generally place high emphasis on leak localization in situations where all the leak locations are known, i.e., under supervised conditions. In practice, it is quite common that partial leak locations are unknown, which are regarded as semi-supervised conditions. In such situations, the OSP issues under semi-supervised conditions ought to be discussed in two aspects. On the one hand, if the data-sets without leak locations account for a small proportion, we just ignore them and the OSP issues can be handled by the abovementioned OSP methods. However, once the data-sets without leak locations are too large to be excluded, the traditional OSP approaches will no longer be applicable.

To solve this problem, this paper develops the semi-supervised OSP problem as a semi-supervised feature selection (FS) problem for deploying the pressure sensors. The monitoring nodes are obtained based on the minimal-Redundancy-Maximal-Relevance (mRMR) criterion [12,13] by applying the semi-JMI criterion [11]. Nevertheless, the traditional semi-supervised FS methods do not take into account the actual scenarios in WDNs, which may lead to the occurrence of blind zones. Therefore, the fuzzy *c*-means (FCM) [14–16] clustering method is adopted to divide the WDN into several zones. Next, we apply the semi-JMI algorithm to each zone. Finally, the proposed semi-supervised OSP strategy monitors the whole WDN by deploying the pressure sensors in the most representative nodes from each zone.

* Corresponding author.
E-mail address: ljuan@jlu.edu.cn (J. Li)

https://doi.org/10.1016/j.jprocont.2019.08.001
0959-1524/

The following contents of this paper are organized as follows: Section 2 introduces the concept of fuzzy *c*-means clustering, and the semi-JMI algorithm is introduced. Section 3 presents the proposed semi-supervised strategy. Section 4 presents the results and the corresponding discussion, and Section 5 concludes the paper.

2. Basic theory

2.1. Fuzzy c-means (FCM) clustering

Fuzzy *c*-means clustering is the most popular fuzzy clustering algorithm. This method can be rendered by minimizing the following objective function:

$$J_m(U,V)=\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ij}^{m}d_{ij}^{2},\quad 1\le m<\infty \tag{1}$$

subject to

$$\sum_{i=1}^{c}u_{ij}=1,\quad 1\le j\le n \tag{2}$$

$$u_{ij}\ge 0,\quad 1\le i\le c,\quad 1\le j\le n \tag{3}$$

$$\sum_{j=1}^{n}u_{ij}>0,\quad 1\le i\le c \tag{4}$$

where $\mathbf{X}=\{x_1,x_2,\ldots,x_n\}\subset\mathbf{R}^s$, s refers to the space dimension; n is the number of samples; m is the fuzzy factor ($m>1$); $d_{ij}=\|x_j-v_i\|$ is the distance between sample x_j and the clustering center v_i, $v_i\subset\mathbf{R}^s$ ($1\le i\le c$); u_{ij} stands for the membership of the jth sample to the ith clustering center; $\mathbf{U}=[u_{ij}]_{n\times c}$ is a matrix of $n\times c$; and $\mathbf{V}=[v_1,v_2,\ldots,v_c]_{c\times s}$ is a matrix of $c\times s$. The steps of the FCM algorithm are as follows [17]:

Step 1: Initialize a membership matrix $\mathbf{U}$ using random numbers, set the fuzzy factor m, choose $\varepsilon>0$ (ε is a small positive constant and is preset to 0.00001 in this paper), and set the iteration counter $k=1$.

Step 2: Calculate the clustering center $\mathbf{V}^k$ using Eq. (5) during the *kth*-step.

$$v_i^k=\frac{\sum_{j=1}^{n}\left(u_{ij}^k\right)^m x_j}{\sum_{j=1}^{n}\left(u_{ij}^k\right)^m}. \tag{5}$$

Step 3: Update $\mathbf{U}^{k+1}$ using Eq. (6).

$$u_{ij}^{k+1}=\frac{1}{\sum_{q=1}^{c}\left(\frac{d_{ij}}{d_{qj}}\right)^{\frac{2}{m-1}}}. \tag{6}$$

Step 4: If $\|\mathbf{U}^{k+1}-\mathbf{U}^{k}\|<\varepsilon$, then stop, we calculate $(J_m)_{\min}$ using Eq. (1); Otherwise, let $k=k+1$ and return to step 2.

2.2. Semi-JMI algorithm

In [18] and later in [19], a different perspective was proposed to minimize the redundancy and maximize the relevancy by applying the joint mutual information (JMI) criterion. It has been proven that the JMI possesses excellent performance in terms of stability and accuracy [20]. The approach focuses on maximizing the complimentary information:

$$J_{jmi}(F_i)=\sum_{F_j\in\gamma}I\left(F_i,F_j;Y\right). \tag{7}$$

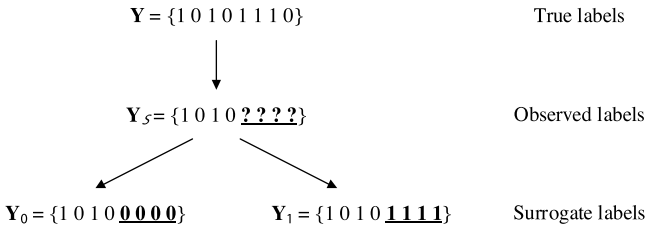

Fig. 1. Surrogate labels of the true label.

Eq. (7) is the information between the targets and a joint random variable (F_i, F_j), which is defined by pairing the candidate F_i with each feature previously selected in γ. The concept is as follows: if the candidate feature F_i is complementary to the existing features, it should be added to γ.

The semi-JMI algorithm based on the JMI criterion is a new method to solve semi-supervised feature selection problems. Usually, a semi-supervised data set D can be divided into two parts: the labeled D_L and the unlabeled D_U. For the convenience of expression, the semi-JMI criterion adopts a simple 1-vs-all strategy to convert the multiclass problems into binary problems; for the labeled D_L, "class 1" remains unchanged, while other labels are set to label "0". Fig. 1 depicts the two patterns the semi-JMI criterion investigates.

Although only some true labels are available, the key to the semi-JMI algorithm is selecting better surrogate labels, i.e., "0" (negative) or "1" (positive), as the "surrogate variable" called "$\mathbf{Y}_0$" and "$\mathbf{Y}_1$", respectively, to replace the blank labels to use for semi-supervised feature selection. As discussed in [11], the false positive rates (FPRs) of the two surrogate tests are the same as the true tests. However, the false negative rates (FNRs) are not (higher than the true tests). This problem can be solved on the premise that the true underlying class probability, $p(y=1)$, is known. We suppose that m stands for the number of positive examples provided, n is the number of negative examples, q represents the number of unlabeled examples, and $p'(y=1)$ is the user's belief in the class probability. Then, the surrogate label can be determined as follows:

$$a=\sqrt{m(m+q)}, \tag{8}$$

$$b=\sqrt{n(n+q)}, \tag{9}$$

$$\psi=\frac{a}{a+b}, \tag{10}$$

where ψ represents the "switching threshold". If $p'(y=1)<\psi$, we use $\mathbf{Y}_0$ as the surrogate label; otherwise, we use $\mathbf{Y}_1$. Finally, the dataset $\mathbf{D}$ is analyzed as a fully labeled dataset by dint of the traditional JMI criterion.

2.3. Support vector machine

The Support Vector Machine (SVM) method [30,31] is applied to identify the corresponding leakage positions (nodes) by inputting the pressure sensitivity matrix $\mathbf{S}$ [22]. We assume that n is the number of samples, given a training set $\{(\mathbf{x}_i,y_i),i=1,2,\ldots,n\}$, and N is the dimensionality of $\mathbf{x}_i$, then the SVM can be represented by the following optimization problem:

$$\min_{w,b,\xi}\left(\frac{1}{2}w^{\mathrm{T}}w+C\sum_{i=1}^{n}\xi_i\right), \tag{11}$$

such that

$$y_i\left(w^T\phi(\mathbf{x}_i)+b\right)\ge 1-\xi_i,\ \xi_i\ge 0,\ \forall i=1,\ldots,n, \tag{12}$$

where C is the penalty parameter; ϕ is the mapping from the training instance space $\mathbf{R}^N$ into a high dimensional feature space; w and b parameterize a hyperplane in the feature space; and ξ stands for a

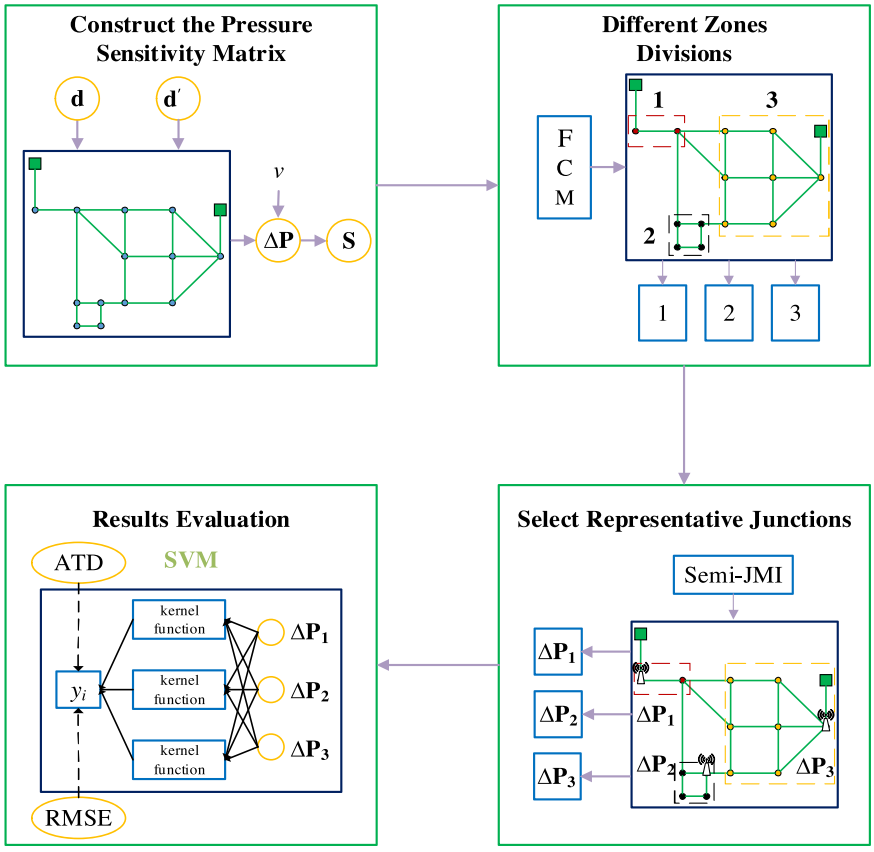

Fig. 2. Leak localization scheme.

slack vector. To solve Eqs. (11) and (12), the above problem is usually transformed into the following quadratic programming (QP) problem:

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j), \tag{13}$$

such that

$$\sum_{i=1}^{n} \alpha_i y_i = 0,\ 0 \le \alpha_i \le C,\ i = 1, 2, \ldots, n \tag{14}$$

where α_i stands for a corresponding Lagrange coefficient. $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^{\mathrm{T}} \phi(\mathbf{x}_j)^{\mathrm{T}}$ is the kernel function. A common and very effective kernel, which is applied in this paper, is the radial basis function (RBF) kernel defined by

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \left| \mathbf{x}_i - \mathbf{x}_j \right|^2), \quad \gamma > 0. \tag{15}$$

For the work of this paper, the software applied to train the SVM is the LIBSVM software [29] of the National Taiwan University.

3. The overall framework

The semi-supervised OSP problem is considered to be a process of semisupervised feature selection. The process of selecting three monitoring nodes to monitor the WDN (predicted leak location y_i) is depicted in Fig. 2.

3.1. Construct the pressure sensitivity matrix

Generally, as long as the nodal demands change, the nodal pressure will change. However, the nodal demands in practice are not always available because there is no need to measure them. The nodal demands can be estimated on the basis of the total demand of the WDN and the statistical ratio of all the nodes. We input the estimated nodal demands (d_1, d_2,..., ds) and other necessary parameters (for instance, the node's elevation and the reservoir pressure and flows) into the EPANET simulator [21] and obtain the normal pressure $\mathbf{P}^0 = \{P_1^0, P_2^0, \ldots, P_s^0\}$.

In the EPANET software, the leakage can be simulated by setting a positive emit coefficient at the nodes or adding additional nodal demands. For convenience, the latter method is adopted because it is easier to accurately control the magnitude of the leakage. After adding the extra nodal demands to one node a time, we express the new nodal demands as $\left(d_1', d_2', \ldots, d_s'\right)$. In addition, for each leak location, the leak magnitudes are within a given range. Supposing that there are s junctions in the WDN, we change the leak nodes and the corresponding leak sizes. Then, the abnormal pressures can be defined as follows:

$$\mathbf{P}^L = \begin{bmatrix} P_{11}^1 P_{21}^1 \cdots P_{s1}^1 \\ P_{12}^1 P_{22}^1 \cdots P_{s2}^1 \\ \vdots \ \ddots \ \vdots \\ P_{1k}^1 P_{2k}^1 \cdots P_{sk}^1 \\ P_{11}^2 P_{21}^2 \cdots P_{s1}^2 \\ \vdots \ \ddots \ \vdots \\ P_{1k}^s P_{2k}^s \cdots P_{sk}^s \end{bmatrix}, \tag{16}$$

where P_{ij}^t indicates the pressure of the i-th node when the j-th leak grade occurs at the t-th node; the leak size grades are from

0 to k, where 0 means no leakage and k represents the maximum leakage. Therefore, the differences between the normal pressures and the abnormal pressures are calculated. Subsequently, the pressure sensitivity matrix **S** is obtained by adding Gaussian noise with the maximal amplitude of 5% from the average of all the pressure residuals, which is noted as follows:

$$\mathbf{S}=\begin{bmatrix} \frac{P_{11}^1-P_1^0+v}{\Delta D_1^1} & \frac{P_{21}^1-P_2^0+v}{\Delta D_1^1} & \cdots & \frac{P_{s1}^1-P_s^0+v}{\Delta D_1^1} \\ \frac{P_{12}^1-P_1^0+v}{\Delta D_2^1} & \frac{P_{22}^1-P_2^0+v}{\Delta D_2^1} & \cdots & \frac{P_{s2}^1-P_s^0+v}{\Delta D_2^1} \\ \vdots & & \ddots & \vdots \\ \frac{P_{1k}^1-P_1^0+v}{\Delta D_k^1} & \frac{P_{2k}^1-P_2^0+v}{\Delta D_k^1} & \cdots & \frac{P_{sk}^1-P_s^0+v}{\Delta D_k^1} \\ \frac{P_{11}^2-P_1^0+v}{\Delta D_1^2} & \frac{P_{21}^2-P_2^0+v}{\Delta D_1^2} & \cdots & \frac{P_{s1}^2-P_s^0+v}{\Delta D_1^2} \\ \vdots & & \ddots & \vdots \\ \frac{P_{1k}^s-P_1^0+v}{\Delta D_k^s} & \frac{P_{2k}^s-P_2^0+v}{\Delta D_k^s} & \cdots & \frac{P_{sk}^s-P_s^0+v}{\Delta D_k^s} \end{bmatrix}, \tag{17}$$

in which v is the random Gaussian noise and ΔD_j^t is the extra nodal demand at node t with the j-th leak grade.

3.2. Divide the WDN into several zones

When somewhere in the WDN leaks, the pressure changes of the adjacent nodes will be within a similar range. In this case, this characteristic can be applied to divide the network into different zones. In [23], k-means clustering [24–26] is applied to divide the WDN into multiple zones. However, in comparison with k-means clustering, the FCM algorithm introduces the concept of membership, whose clustering results are theoretically more ideal. Therefore, the FCM clustering deals with the first s columns of **S,** and each column will be assigned to c clustering centers. Specifically, the network is divided into c zones.

3.3. Semi-supervised feature selection

After clustering the different nodes, the semi-JMI criterion is utilized to select the representative nodes to deploy the pressure sensors. Its implementation details are as follows:

(1) Transform the supervised problems into semi-supervised problems.

To simulate the true loss of the leak location labels or partial locations that have not been recorded in the WDN, this paper randomly removes 75% of all the labels.

(2) Making the surrogate labels.

Since 75% of the leak location labels have been removed, surrogate labels are necessary to compose the label sets to guarantee improvements in solving the feature selection problems [11]. As discussed in Section 2.2, in the remaining labels, we assume that one class of the remaining labels (take class 1 as an example) is positive ("1"), and the others are negative ("0"). Then, the class probability is $p'(y=1)=1/s$. If we switch the threshold $\psi > 1/s$, then the missing labels are "0"; otherwise, the missing labels are "1".

(3) Select the representative nodes.

After grouping similar columns of **S** and making the surrogate labels, we rank the features (columns of **S**) by applying the JMI criterion. The features with the minimum redundancy and maximum relevance in each group are selected, and their corresponding nodes constitute the representative nodes.

3.4. Results evaluation

It can be seen from the experiment that under the same leakage condition (leak in the same node with the same leak grade) in the WDN, the pressure changes of the adjacent nodes are almost similar. In addition, there is no difference in the pressure changes in the nodes closest to the reservoir in the EPANET software, regardless of whether other nodes are leaking at the same leak magnitude. These two aspects greatly reduce the accuracy of the leak localization classification on the basis of the machine learning algorithms. In general, it is not wise to assess the accuracy of the leak localization results. However, lower accuracy does not mean that machine learning methods are not available in this case. Although their results always fail to correctly predict the actual leak locations, they are always close to the actual leak nodes. In this way, the prediction will play a part in locating the real leak nodes.

An effective indicator of the evaluation results, called the average topological distance (ATD), is presented in [10]. It represents the average topological distance between the predicted locations and the actual leaking locations. The ATD is calculated as

$$ATD=\frac{\sum_{i=1}^{n_c}\sum_{j=1}^{n_c}N_{i,j}D_{i,j}}{\sum_{i=1}^{n_c}\sum_{j=i}^{n_c}N_{i,j}}, \tag{18}$$

where n_c refers to the number of features (nodes); $N_{i,j}$ represents the number of times; the leak node i is considered to be node j, and $D_{i,j}$ is the topological distance (the number of pipes in the paper) between node i and node j.

The root mean square error (RMSE) is regularly employed to evaluate the performance of the results. This metric explains the stability of the predicted results. For the sake of simplicity, we assume that e_i ($i=1, 2, \ldots, n$) represents the errors of n samples, and the RMSE is calculated for the dataset as

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}e_i^2}. \tag{19}$$

To evaluate the deviation of the average topological distance between the predicted nodes and actual nodes, this paper defines the RMSE of the ATD by regarding $D_{i,j}$ as the errors (e_i) of the ATD:

$$RMSE_{ATD}=\sqrt{\frac{\sum_{i=1}^{n_c}\sum_{j=1}^{n_c}N_{i,j}D_{i,j}^2}{\sum_{i=1}^{n_c}\sum_{j=1}^{n_c}N_{i,j}}}. \tag{20}$$

The predicted locations and their actual leak locations are obtained by inputting the pressure sensitivity matrix, **S** (after feature selection). Accordingly, the corresponding leak locations (l_i labels) for the SVM, $D_{i,j}$ are calculated by comparing the error of their topological distance in the WDN. On this basis, the ATD and its RMSE will be resolved.

4. Case studies

In this section, the original semi-JMI criterion and the proposed semi-supervised strategy are applied to two WDNs (a simple network from [27] and a more complicated WDN from [28], respectively). The EPANET driving program is implemented to simulate the actual WDNs. The National Taiwan University's LIBSVM software is applied to classify the predicted leak locations. The ATD and its RMSE are calculated based on the SVM results. In addition, two

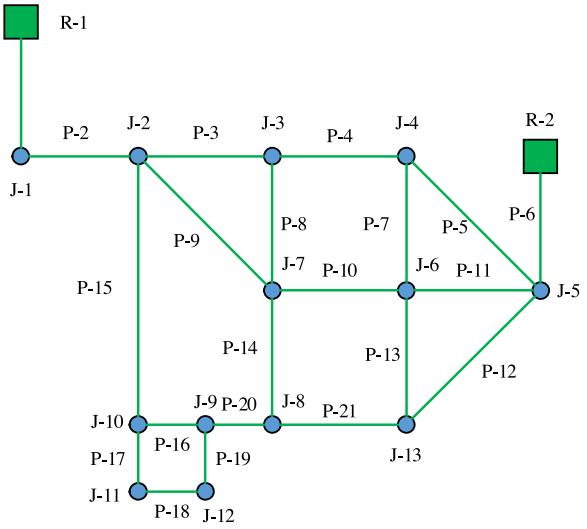

Fig. 3. Layout of Network 1.

Table 1
Nodal properties of Network 1.

Node ID	Elevation (m)	Required demand (L/s)
J-1	27.43	0.00
J-2	33.53	59.0
J-3	28.96	59.0
J-4	32.00	178.0
J-5	30.48	59.0
J-6	31.39	190.0
J-7	29.56	178.0
J-8	31.39	91.0
J-9	32.61	0.00
J-10	34.14	0.00
J-11	35.05	30.0
J-12	36.58	30.0
J-13	33.53	0.00
R-1	60.96	N/A
R-2	60.96	N/A

semi-supervised methods are compared. For the computing time, all the calculations were performed using a desktop computer with an INTEL CORE i5-8400U CPU @ 2.80 [GHz], 8 [GB] of RAM memory as well as a Windows 10 Home 64 bits OS. Meanwhile, the MATLAB 2016a software is applied.

4.1. Case study 1

A simple network is shown in Fig. 3. The network consists of 2 reservoirs, 13 nodes, and 21 pipes. The nodal properties of Network 1 are given in Table 1; the pipe characteristics are listed in Table 2; and the daily average demand is approximately 874 L/s.

Although there are no existing criteria with regard to the leak magnitudes, it is significant to ensure that the abnormal pressure values caused by leakage are different from the normal pressure values. In addition, it should be noted that the maximum leakage discharge cannot be set too large. Otherwise, the WDN will not work in the EPANET software. In this case, the amount of the water flow to the surface simulation is more or less than 3% of the average demand, which ranges from 19 L/s to 37 L/s at intervals of 2 L/s added to the selected nodes, generating 130 leak cases. The input of our method is the pressure sensitivity matrix with a size of 13×130. Subsequently, the WDN is divided into several zones according to the FCM algorithm. As the fuzzy factor m increases, the clustering number increases from 1 to 13. Regardless of how many zones are divided into Network 1, it is convenient to obtain the clustering results. In practice, the number of zones depends on the actual

Table 2
Pipe characteristics of Network 1.

Pipe ID	L (m)	D (mm)	C (HW)
P-1	609.60	762	130
P-2	243.80	762	128
P-3	1524.00	609	126
P-4	1127.76	609	124
P-5	1188.72	406	122
P-6	640.08	406	120
P-7	762.00	254	118
P-8	944.88	254	116
P-9	1676.40	381	114
P-10	883.92	305	112
P-11	883.92	305	110
P-12	1371.60	381	108
P-13	762.00	254	106
P-14	822.96	254	104
P-15	944.88	305	102
P-16	579.00	305	100
P-17	487.68	203	98
P-18	457.20	152	96
P-19	502.92	203	94
P-20	883.92	203	92
P-21	944.88	305	90

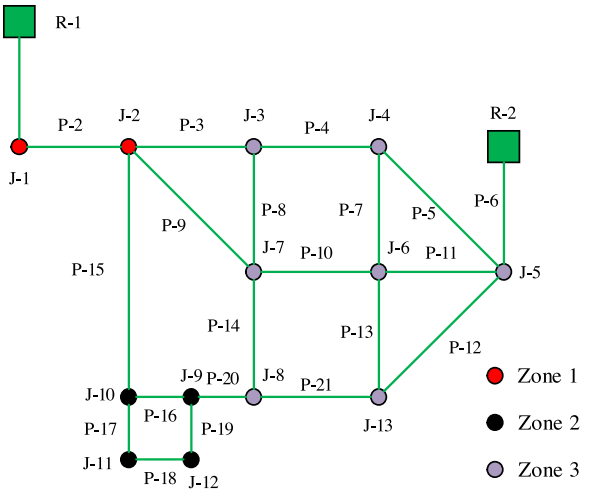

Fig. 4. Three zones division in Network 1.

Table 3
Results of two semi-supervised OSP strategies in Network 1.

Data set	Method 1	Time (s)	Method 2	Time (s)
1	1,2,5	0.123	2,3,9	0.303
2	2,4,5	0.125	2,4,9	0.305
3	2,4,5	0.120	2,4,9	0.300
4	1,2,4	0.124	2,4,9	0.304
5	1,4,5	0.123	1,4,9	0.303
6	2,4,5	0.123	2,4,9	0.303
7	1,2,5	0.124	2,4,9	0.304
8	2,4,5	0.125	2,4,9	0.305
9	1,4,5	0.123	1,4,9	0.303
10	2,4,5	0.124	2,4,9	0.304
Average	2,4,5	0.123	2,4,9	0.303

requirement and budget constraints. In this paper, Network 1 in Fig. 3 is divided into 3 zones (see Fig. 4).

As shown in Fig. 4, Network 1 is divided into 3 zones, which are composed of 2, 4 and 7 nodes. Each zone consists of several adjacent nodes with similar pressure change rules. The ultimate goal of the OSP problem is to deploy the pressure sensors in the most representative nodes. Two semi-supervised methods are applied to select the monitoring nodes. The results of the two strategies are summarized in Table 3.

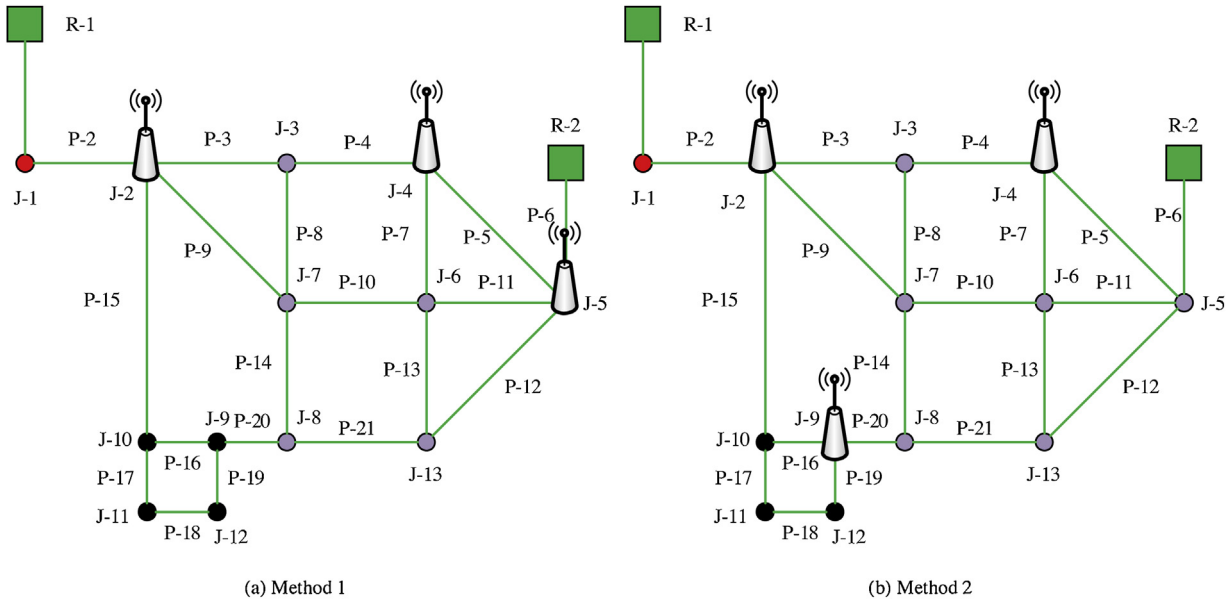

Fig. 5. Different OSP strategies of two methods in Network 1.

As shown in Table 3, the monitoring nodes and the computing time are different each time according to the two semi-supervised methods. Method 1 represents the semi-JMI algorithm as well as the proposed strategy for Method 2. It can be explained that the leak location labels are deleted randomly when transforming the fully supervised data into semi-supervised data. Due to the process of FCM clustering, the computing time of Method 2 is 0.180 s longer than that of Method 1. Different data sets produce different combinations of nodes. The final selected nodes are the set {2, 4, 5} for Method 1 and {2, 4, 9} for Method 2, as depicted in Fig. 5. The different node sets are selected based on the two methods presented in Fig. 5. The two methods are the same for {2, 4}, but the third node is not. Method 1 selects node 5 to belong to the same zone as node 4, regardless of the monitoring of Zone 2. In contrast, node 9 is selected in Method 2. The monitoring pressure sensors in Method 2 are distributed more discretely in all the zones relative to the distribution of Method 1 due to the FCM pretreatment. Method 2 can monitor all the leak cases, which is more accurate than Method 1. Method 1 only takes into account the relevance and redundancy without the actual situation of the WDN monitoring. Meanwhile, Method 2 also contains the pressure changes for all the different zones apart from the relevance and redundancy. Therefore, the results show that Method 2, rather than Method 1, is a better choice for dealing with OSP problems in WDNs. According to the two semi-supervised methods, the unselected pressure changes are removed. Subsequently, for each method, 130 cases are amalgamated; 87 cases are randomly selected to form the training set, and the remaining 43 cases are applied as the validation set in LIBSVM software. The results of the SVM are not very stable due to the different training cases. To overcome this problem, 87 cases are randomly selected from 130 cases 10 times. The ATD and its RMSE are applied to evaluate the performance of the two methods, and are summarized in Tables 4 and 5.

As seen from Tables 4 and 5, the SVM results on the basis of Method 1 accurately predict the leak locations twice, while Method 2 predicts five times. The ATD of Method 1 is 3.340 times more than that of Method 2, and the RMSE of Method 1 is 2.109 times more than that of Method 2. This means that, on the basis of Method 1, the farther the average distance is between the predicted locations and actual positions, the less stable the predicted location is.

Table 4
The results of ATD and RMSE based on Method 1 in Network 1.

Data set	ATD	RMSE
1	0.395	0.699
2	0.488	0.762
3	0.628	0.952
4	0.465	0.778
5	0.791	1.078
6	0.535	0.849
7	0.488	0.849
8	0.419	0.747
9	0.419	0.778
10	0.581	0.835
Average	0.521	0.833

Table 5
The results of ATD and RMSE based on Method 2 in Network 1.

Data set	ATD	RMSE
1	0.093	0.305
2	0.140	0.431
3	0.070	0.264
4	0.256	0.665
5	0.279	0.571
6	0.256	0.431
7	0.186	0.482
8	0.116	0.341
9	0.163	0.457
10	0	0
Average	0.156	0.395

4.2. Case study 2

Fig. 6 displays the network layout of the EPANET Net3, including 92 nodes, 117 pipes, two reservoirs (one lake and one river), three tanks, and two pumps. Furthermore, on the one hand, the nodes {10, 15, 35, 123, 203} are not consumer nodes; on the other hand, nodes {20, 40, 50, 60, 61, 601} are too close to reservoirs, tanks, or pumps. These 10 nodes are not considered to be leakage nodes or sensor placement nodes. The daily average flow of this network is approximately 3048.11 L/s. The hourly coefficients of the nodal demands for the remaining 81 nodes in EPANET Net3 are depicted in Fig. 7. The amount of water flow to the surface simulation is set between 1.97% and 5.91% of the daily average flow of the network,

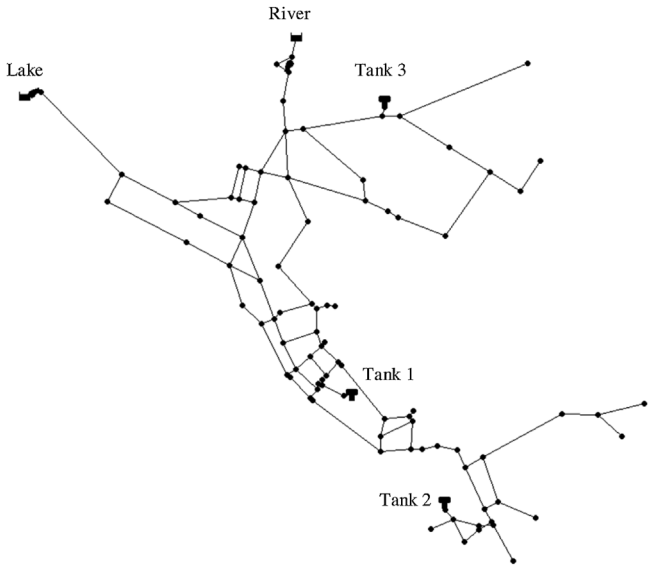

Fig. 6. Layout of the EPANET Net3.

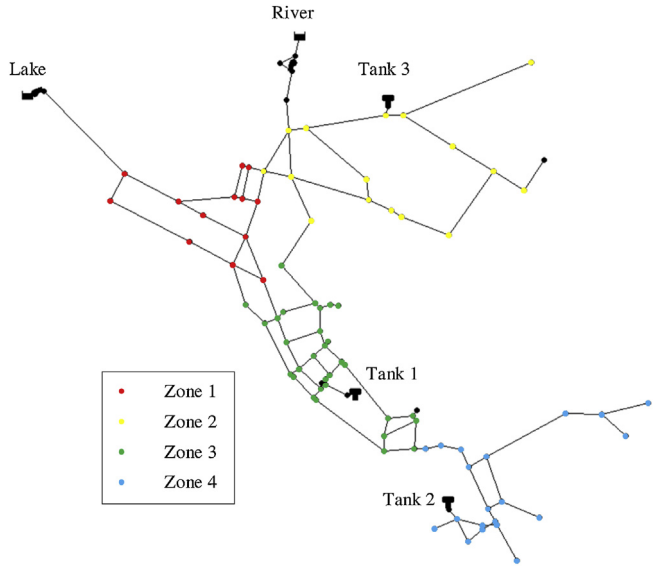

Fig. 8. The results of FCM clustering in the EPANET Net3.

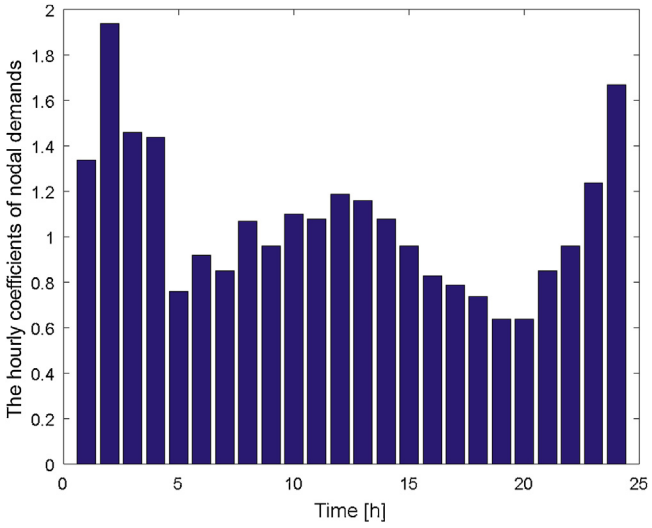

Fig. 7. The hourly coefficients of the nodal demands in the EPANET Net3.

with the amount added to the selected nodes varying from 60 L/s to 180 L/s. The time step Δt is 1 h, and the flow step Δq is 30 L/s. Consequently, the number of leakage cases reached 9720(8124*5).

The EPANET Net3 is divided into 4 zones by applying FCM clustering preprocessing. The input data in this process is a matrix of 81×9720, which is the pressure sensitivity matrix of the network. The different rows of the matrix are divided into artificial sets that are preset to stand for different zones. The adjacent nodes are assigned to the same zone in Fig. 8. The pressure sensitivities of adjacent nodes are similar in value.

The next step is to select the representative nodes from each zone. All the nodes have *Nc* combinations as follows:

$$Nc = \frac{n_f!}{n_s!(n_f - n_s)!}, \tag{21}$$

where n_f refers to the number of all the nodes, and n_s refers to the number of nodes to select. In general, it is not easy to select n_s nodes from n_f nodes. For instance, in the EPANET Net3, n_f = 81, n_s = 4, then N_c = 1,663,740. However, only a small number of all the combinations are suitable for predicting the leak locations, and the semi-supervised strategy intends to select the representative nodes to monitor the entire network. As described in Case Study 1, the FCM algorithm clustering is applied to cluster the nodes that belong to the same zone, and then the semi-JMI algorithm is applied to select the monitoring nodes from each zone. The results of the proposed semisupervised method with 4 monitoring nodes are summarized in Table 6. The average computing time is 13.46 s, including the FCM clustering pretreatment time of 0.98 s. In this paper, we adopt the fuzzy c-means (FCM) clustering as well as a semi-JMI algorithm integrated method to select the monitoring nodes of the entire WDN. Furthermore, our method is under semi-supervised conditions (75% of the labels are lost or randomly deleted). For the abovementioned reasons, once we determine the number of monitoring nodes, the monitoring nodes that we choose may be different each time. According to the experiments, it can be found that several fixed nodes {109, 119, 184, 229} can appear multiple times. In the remaining experiments, the new nodes may be close to the fixed nodes, which we choose as the monitoring nodes. However, they rarely appear (at most three times). Finally, the selected nodes of the proposed method are the set {109, 119, 184, 229} as depicted in Fig. 9. The traditional semi-JMI method without the FCM clustering pretreatment selects the node set {109, 119, 149, 229}, which is shown in Table 6.

Table 6
Results of the proposed semi-supervised OSP strategy in the EPANET Net3.

Data set	Monitoring nodes	Time (s)
1	109, 119, 184, 229	13.44
2	109, 120, 184, 229	13.46
3	109, 119, 197, 229	13.51
4	109, 119, 184, 229	13.43
5	109, 119, 184, 229	13.40
6	109, 120, 184, 229	13.48
7	109, 119, 197, 229	13.46
8	109, 119, 184, 229	13.42
9	109, 119, 184, 229	13.52
10	109, 119, 197, 229	13.44
Average	109, 119, 184, 229	13.46

For the two strategies, the proposed method selects the representative monitoring nodes from 4 different zones, which avoids the appearance of blind monitoring zones. The traditional semi-JMI algorithm can also select 4 useful monitoring nodes However, as shown in Fig. 10, the semi-JMI method selects 2 monitoring nodes from Zone 2, but none from Zone 3. The results of these two meth-

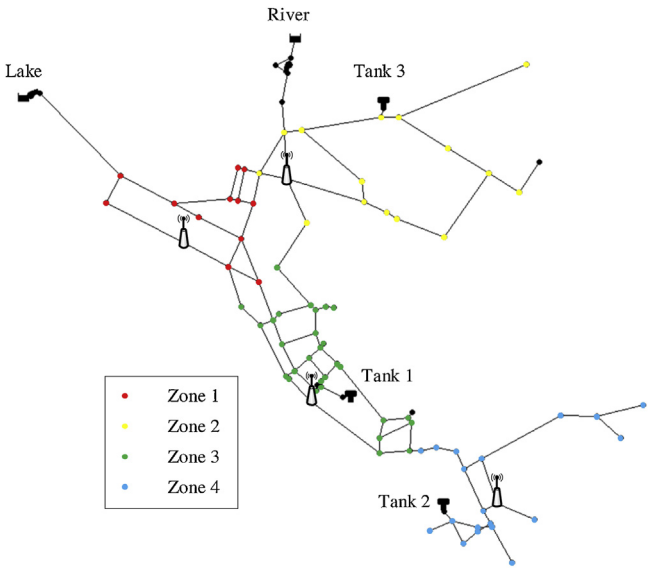

Fig. 9. The monitoring nodes selected by the proposed method in the EPANET Net3.

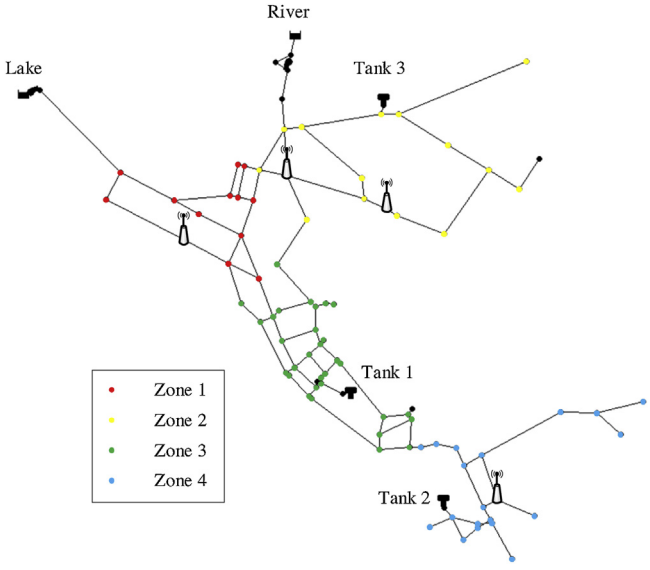

Fig. 10. The monitoring nodes selected by the Semi-JMI algorithm in the EPANET Net3.

Table 7
The results of ATD and RMSE based on semi-JMI method in EPANET Net3.

Data set	ATD	RMSE
1	1.27	1.77
2	1.93	3.10
3	1.80	2.35
4	1.07	1.55
5	1.00	1.39
6	1.27	1.73
7	0.87	1.18
8	1.27	1.88
9	1.80	2.59
10	2.00	3.06
Average	1.43	2.06

ods is reflected in Tables 7 and 8. The ATD of the semi-JMI method is 1.43, and the RMSE is 2.06. The ATD of the proposed method is 1.24, and the RMSE is 1.80. Therefore, the proposed method is proven to be superior to the semi-JMI method in terms of the ATD and RMSE. The first and most important point of this condition is that the monitoring nodes selected by the semi-JMI method cannot cover all 4 zones. When the leakage occurs in Zone 3, the localization results of the monitoring nodes selected by the semi-JMI method are less accurate than those of the proposed method, for which the monitoring nodes locate all 4 zones. However, the EPANET Net3 is more complex than Network 1, resulting in a decrease in the classification accuracy. This causes the ATD and RMSE values to be higher. In addition, the method also deploys 3 and 5 pressure sensors in the EPANET Net3. For 3 monitoring nodes, the proposed method selects nodes {109, 119, 229}. Accordingly, the average ATD and RMSE results are 1.89 and 2.39, respectively. For 5 monitoring nodes, nodes {109, 119, 184, 199, 229} are selected, and the results for average ATD and RMSE are 1.17 and 1.56, respectively. From these results, it can be summarized that the ATD of 3 monitoring nodes decreases by 34.4% compared with that of 4 monitoring nodes, and the RMSE decreases by 32.8%. However, 5 monitoring nodes only increase the ATD by 5.6% and the RMSE by 13.3% than 4 monitoring nodes. Within the allowable error range, 4 monitoring nodes are sufficient to monitor the entire EPANET Net3.

Table 8
The results of ATD and RMSE based on the Method 2 in EPANET Net3.

Data set	ATD	RMSE
1	1.33	1.90
2	0.93	1.26
3	1.47	2.13
4	1.40	2.02
5	1.33	1.86
6	1.40	2.11
7	0.93	1.41
8	0.87	1.24
9	1.40	2.02
10	1.33	2.00
Average	1.24	1.80

5. Conclusion

In this paper, a novel semi-supervised strategy is proposed to deploy pressure sensors to monitor the leakage of a WDN. Moreover, FCM clustering is applied to divide the network into several zones. Then, the representative nodes are selected from each zone by applying the semi-JMI criterion. Two case studies are discussed to illustrate and validate the effectiveness of the proposed method. In Case Study 1, the effectiveness of the method is analyzed from two aspects: the ATD and its RMSE. The results indicate that better results can be obtained by applying the FCM and semi-JMI integrated method in comparison with the traditional semi-JMI method without the FCM pretreatment. In Case Study 2, the proposed method is applied to a more complex WDN. As shown in Case Study 1, the proposed method is better in deployment than semi-JMI alone, which ensures that there is no blind monitoring zone in the WDN. This approach can significantly improve the efficiency and effectiveness of leak detection in WDNs.

Declaration of Competing Interest

The authors declared that they have no conflicts of interest to this work.

Acknowledgements

This paper is supported by the key Science Foundation of the Department of Science and Technology of Jilin Province (Grant No. 20180201081SF, 20190303082SF), Jilin Provincial Special Funding for Industrial Innovation (Grant No. 2017C031-1), National Natural Science Foundation of China (No.61771219) and the Fundamen-

tal Research Funds of Jilin University (No. SXGJQY2017-9 and No. 2017TD-19). Thanks for the permission to publish this paper.

References

[1] G.S. Sankar, S.M. Kumar, S. Narasimhan, S. Narasimhan, M. Bhallamudi, Optimal control of water distribution networks with storage facilities, J. Process Control 32 (2015) 127–137.
[2] Y. Wang, V. Puig, G. Cembrano, Non-linear economic model predictive control of water distribution networks, J. Process Control 56 (2017) 23–34.
[3] R. Puust, Z. Kapelan, D.A. Savic, T. Koppel, A review of methods for leakage management in pipe networks, Urban Water J. 7 (1) (2010) 25–45.
[4] M. Aral, Guan Mustafa, Maslia Jiabao, L. Morris, Optimal design of sensor placement in water distribution networks, J. Water Resour. Plan. Manage. 136 (1) (2010) 5–18.
[5] M.S. Khorshidi, M.R. Nikoo, M. Sadegh, Optimal and objective placement of sensors in water distribution systems using information theory, Water Res. 143 (2018) 218–228.
[6] B. Farley, S.R. Mounce, J.B. Boxall, Field testing of an optimal sensor placement methodology for event detection in an urban water distribution network, Urban Water J. 7 (6) (2010) 12.
[7] R. Sarrate, J. Blesa, F. Nejjari, Clustering techniques applied to sensor placement for leak detection and location in water distribution networks, in: 2014 22nd Mediterranean Conference of Control and Automation (MED), 2014, pp. 109–114.
[8] J. Blesa, F. Nejjari, R. Sarrate, Robust sensor placement for leak location: analysis and design, J. Hydroinform. 18 (1) (2016) 136–148.
[9] A. Soldevila, R.M. Fernandez-Canti, J. Blesa, S. Tornil-Sin, V. Puig, Leak localization in water distribution networks using Bayesian classifiers, J. Process Control 55 (2017) 1–9.
[10] A. Soldevila, J. Blesa, S. Tornil-Sin, R.M. Fernandez-Canti, V. Puig, Sensor placement for classifier-based leak localization in water distribution networks using hybrid feature selection, Comput. Chem. Eng. 108 (2018) 152–162.
[11] K. Sechidis, G. Brown, Simple strategies for semi-supervised feature selection, Mach. Learn. 107 (5) (2017) 1–39.
[12] H. Peng, F. Long, C. Ding, Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (8) (2005) 1226–1238.
[13] P. Maji, P. Garai, On fuzzy-rough attribute selection: criteria of max-dependency, max-relevance, minredun-ancy, and max-significance, Appl. Soft Comput. 13 (9) (2013) 3968–3980.
[14] K.S. Tan, H.L. Wei, N.A.M. Isa, Novel initialization scheme for fuzzy c-means algorithm on color image segmentation, Appl. Soft Comput. 13 (4) (2013) 1832–1852.
[15] S.K. Adhikari, J.K. Sing, D.K. Basu, M. Nasipuri, Conditional spatial fuzzy c-means clustering algorithm for segmentation of mri images, Appl. Soft Comput. 34 (2015) 758–769.
[16] Z. Ji, Y. Xia, Q. Chen, Q. Sun, D. Xia, D.D. Feng, Fuzzy c-means clustering with weighted image patch for image segmentation, Appl. Soft Comput. 12 (6) (2012) 1659–1667.
[17] S. He, N. Belacel, H. Hamam, Y. Bouslimani, Y. Fuzzy, Clustering with improved artificial fish swarm algorithm, International Joint Conference on Computational Sciences and Optimization, IEEE 2 (2009) 317–321.
[18] H.H. Yang, J. Moody, Data visualization and feature selection: new algorithms for nongaussian data, Adv. Neural Inf. Process. Syst. 12 (2009).
[19] F. Meyer, D. Paarmann, M. D'Souza, R. Olson, E.M. Glass, M. Kubal, et al., The metagenomics rast server-a public resource for the automatic phylogenetic and functional analysis of metagenomes, BMC Bioinform. 9 (1) (2008) 386.
[20] G. Brown, A. Pocock, M.J. Zhao, M. Luján, Conditional likelihood maximisation: a unifying framework for information theoretic feature selection, J. Mach. Learn. Res. 13 (1) (2012) 27–66.
[21] L.A. Rossman, EPANET 2 user's manual, EPA/600/R-00/057, U.S.EPA, Cincinnati, 2000.
[22] Ramon Pérez, V. Puig, J. Pascual, J. Quevedo, E. Landeros, A. Peralta, Methodology for leakage isolation using pressure sensitivity analysis in water distribution networks, Control Eng. Pract. 19 (10) (2011) 1157–1167.
[23] Q. Zhang, Z.Y. Wu, M. Zhao, J. Qi, Y. Huang, H. Zhao, Leakage zone identification in large-scale water distribution systems using multiclass support vector machines, J. Water Resour. Plan. Manage. 142 (11) (2016) 1–15.
[24] J. Liu, Q. Li, W. Chen, T. Cao, A discrete hidden Markov model fault diagnosis strategy based on k-means clustering dedicated to pem fuel cell systems of tramways, Int. J. Hydrogen Energy 43 (27) (2018).
[25] M.E. Celebi, H.A. Kingravi, P.A. Vela, A comparative study of efficient initialization methods for the k-means clustering algorithm, Expert Syst. Appl. 40 (1) (2012) 200–210.
[26] T. Kanungo, D.M. Mount, N.S. Netanyahu, C.D. Piatko, R. Silverman, A.Y. Wu, An efficient k-means clustering algorithm: analysis and implementation, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 881–892.
[27] K.N. Mallick, I. Ahmed, K.S. Tickle, K.E. Lansey, Determining pipe groupings for water distribution networks, J. Water Resour. Plan. Manage. 128 (2) (2002) 130–139.
[28] USEPA, EPANET 2.0, Available from:, 2002 http://www2.epa.gov/water-research/epanet.
[29] C.C. Chang, C.J. Lin, LIBSVM: A Library for Support Vector Machines, ACM, 2011.
[30] S. Bouhouche, L.L. Yazid, S. Hocine, Jürgen Bast, Evaluation using online support-vector-machines and fuzzy reasoning. Application to condition monitoring of speeds rolling process, Control Eng. Pract. 18 (9) (2010) 1060–1068.
[31] X. Xiong, L. Chen, J. Liang, A new framework of vehicle collision prediction by combining SVM and HMM, IEEE Trans. Intell. Transp. Syst. PP(99 (2018) 1–12.