Memorias del Congreso Nacional de Control Automático
ISSN: 2594-2492

Localización de fugas en redes de distribución de agua mediante k-NN con distancia cosenoidal

Ildeberto Santos-Ruiz [*,1] **F. R. López-Estrada** [*] **Vicenç Puig** [**]
Joaquim Blesa [**] **Mohammadreza Javadiha** [**]

[*] *Tecnológico Nacional de México*
Instituto Tecnológico de Tuxtla Gutiérrez
TURIX-Dynamics Diagnosis and Control Group
Carretera Panamericana km 1080 S/N,
29050, Tuxtla Gutiérrez, México.

[**] *Universitat Politècnica de Catalunya*
Institut de Robòtica i Informàtica Industrial, CSIC-UPC.
Parc Tecnològic de Barcelona. C/ Llorens i Artigas 4-6,
08028, Barcelona, España.

Resumen: Se propone la localización de fugas en redes de distribución de agua mediante clasificadores basados en el método de los vecinos más cercanos (k-NN) con métrica de distancia cosenoidal. El uso de distancias cosenoidales mejora la respuesta del clasificador, con relación al que usa métrica Euclidiana. Comparado con las técnicas de localización de fugas basadas en la máxima correlación de los residuos, se consigue una mayor robustez en condiciones altamente ruidosas, y una menor dependencia del modelo hidráulico de la red, lo que facilita su implementación, pues no requiere del cálculo de la matriz de sensibilidad. La técnica propuesta se programó en MATLAB® y se probó con datos sintéticos obtenidos de simulaciones con EPANET. La evaluación del desempeño reportada se basa en el índice de pérdidas (la fracción de fugas localizadas erróneamente) y en una medida del error de localización obtenida de la distancia topológica.

Palabras clave: Localización de fugas, redes de distribución de agua, diagnóstico de fallas, métodos basados en datos, k-NN, distancia cosenoidal, distancia de Mahalanobis.

1. INTRODUCCIÓN

La pérdida por fugas en los sistemas de distribución es uno de los principales problemas al gestionar el agua potable en los núcleos urbanos. El porcentaje de agua químicamente tratada que se pierde en las tuberías antes de llegar a los consumidores finales ronda el 30 % a nivel mundial, y en algunas ciudades de México supera el 60 % (OECD, 2016). Las fallas por rotura en una tubería, incluso pequeñas, generan la pérdida de grandes volúmenes de agua cuando se mantienen sin reparar durante mucho tiempo, por lo que es necesario detectarlas en un corto plazo y localizarlas con una exactitud que permita su rápida reparación.

Debido a que las fugas no siempre son visibles, porque el agua fugada puede filtrarse hacia abajo de la tubería en lugar de emerger hacia la superficie, la localización precisa de una fuga podría requerir el uso de instrumentación especializada (medidores de vibración, geófonos, etc.), además de sistemas computacionales que mediante el monitoreo de las variables hidráulicas de la red (presiones y caudales) permitan alertar sobre la existencia de fugas y acotarlas en una zona de pocos metros para facilitar el trabajo del personal de mantenimiento. Puig et al. (2017) describen las principales técnicas de monitoreo y diagnóstico en tuberías y redes de distribución de agua, así como las estrategias de control usadas frecuentemente para minimizar el efecto de las fugas.

La localización de fugas en redes de agua puede abordarse como un problema de detección y aislamiento de fallas (FDI, por sus siglas en inglés), y en la literatura se reportan diferentes métodos que consideran tanto enfoques basados en modelos (*model-based*) como otros basados en datos (*data-driven*). Existen diferentes métodos de FDI basados en modelos que han demostrado su utilidad en diversas áreas, pero no son aplicables directamente a

[1] Autor corresponsal: idelossantos@ittg.edu.mx.
Está investigación fué financiada por el Consejo Nacional de Ciencia y Tecnológia (CONACYT) a través de la convocatoria Atención a Problemas Nacionales, con núm. de proyecto PN-2016/3595.

la localización de fugas debido a que las ecuaciones del modelo hidráulico de las redes no son explícitas, por lo que deben usarse métodos numéricos para resolverlas en cada instante de tiempo, dificultando su aplicación. Esto ha motivado el desarrollo de métodos de FDI específicos para la localización de fugas, como los propuestos por Pérez et al. (2011), que se basan en mediciones presión en algunos nodos de la red y en un análisis de sensibilidad a las fugas de esas presiones. En esa metodología, se calculan continuamente los residuos (diferencia entre las presiones instantáneas y las presiones nominales sin fuga, estimadas con un modelo hidráulico de la red) y se comparan contra umbrales que se establecen en función del ruido y la incertidumbre del modelo; cuando algún residuo sobrepasa su umbral (lo que indica una fuga) se consulta la matriz de sensibilidad para determinar cuál fuga está presente. Casillas et al. (2013) han propuesto algunas mejoras en la técnica básica de localización de fugas mediante residuos considerando un horizonte de tiempo. La aplicación de la metodología basada en residuos usando la matriz de sensibilidad en una red de distribución real (en Barcelona, España) ha sido descrita por Perez et al. (2014), donde el procedimiento consiste en medir la correlación entre el vector de residuos de la fuga desconocida y cada uno de los vectores en la matriz de sensibilidad, buscando la máxima similitud que hipotéticamente corresponde al nodo con fuga. Geométricamente, la máxima correlación se presenta cuando el vector de residuos se orienta en la misma dirección que uno de los vectores de la matriz de sensibilidad. La hipótesis subyacente en esta metodología es que los vectores de residuos correspondientes a fugas en el mismo nodo siguen más o menos la misma dirección (los mismos ángulos en el espacio de los residuos), independiente de su magnitud.

Una característica del método de localización de fugas basado en las direcciones de los residuos, es que su desempeño se degrada cuando las mediciones de presión son ruidosas, pues un pequeño cambio en cualquier componente del vector de residuos hace que la dirección de este cambie significativamente, y así podría apuntar en la dirección de las fugas correspondientes a otro nodo, conduciendo a una falsa ubicación. Para desarrollar una técnica de localización más robusta se propone un sistema que no solo considere la dirección característica más cercana al vector de residuos bajo prueba (el correspondiente a la fuga actual) sino un subconjunto que incluya las k direcciones más cercanas ($k > 1$). Como se describirá en la siguiente sección, esta formulación corresponde a un clasificador k-NN (siglas en inglés de k-Nearest Neighbors) bajo la métrica no Euclidiana denominada "distancia cosenoidal". A diferencia del clasificador k-NN Euclidiano, donde las muestras de la misma clase se agrupan en regiones con geometría esférica, al usar k-NN con distancia cosenoidal las diferentes muestras de fuga forman clases por similitud de orientación (ángulo), agrupándose en regiones más bien cónicas. Hipotéticamente, esto dificultaría clasificar las fugas por su magnitud, pero no por su ubicación, que es el propósito del estudio. En este trabajo también se analizará la métrica no Euclidiana denominada "distancia de Mahalanobis"; al final se evaluará el desempeño de ambas y se compararán con los resultados obtenidos con el método de la matriz de sensibilidad. Ferrandez-Gamot et al. (2015) y Soldevila et al. (2016) han abordado antes la localización de fugas mediante clasificadores k-NN, pero usando el concepto Euclidiano de distancia, por lo que ahora se extiende esa línea de investigación.

La propuesta que se presenta se enmarca dentro de las técnicas para localización de fugas mediante el monitoreo de presiones en estado estable, pues ignora la dinámica del proceso, pero es compatible con la variación en el consumo de los usuarios a lo largo del día. La razón para que esta propuesta se base en mediciones de presión es que estas son más sensibles a las fugas que las mediciones de caudal, además de que los sensores de presión son menos costosos y más fáciles de instalar y mantener. Una consecuencia de la alta sensibilidad en las presiones es que estas también son más susceptibles al ruido, por lo que es necesario seleccionar cuidadosamente los nodos donde se colocarán los sensores, para maximizar la captura de información relevante sobre las fugas y a la vez minimizar el efecto del ruido. Sin embargo, para delimitar este trabajo se asumirá que la ubicación de los sensores de presión se ha predeterminado, de modo que su colocación óptima no es objeto de estudio. Se pueden consultar algunas propuestas para la colocación óptima de sensores en los trabajos de Casillas et al. (2015) y Blesa et al. (2016).

El resto del documento se organiza de la siguiente manera: En la Sección 2 se presenta el fundamento matemático de la clasificación k-NN con diferentes métricas de distancia, y se describe su aplicación a la localización de fugas. En la Sección 3 se presentan los resultados de aplicar el método propuesto con datos de fuga de la red de Hanoi (Vietnam), y se compara su desempeño usando diferentes métricas de distancia. Finalmente, en la Sección 4 se presentan las conclusiones y la perspectiva de futuros trabajos en la investigación sobre localización de fugas.

2. METODOLOGÍA

El algoritmo k-NN es una extensión del método del vecino más cercano (NN, *Nearest Neighbor*). En este método se recolecta una serie de muestras etiquetadas que contienen mediciones de algunas características (*features*) de los objetos o eventos que se desea clasificar. Luego, para clasificar nuevas muestras se usan medidas de distancia para identificar cuál es la muestra del conjunto inicial, denominado conjunto de entrenamiento, que más se parece a la muestra bajo prueba (la más cercana, en el espacio de características), y se asume que dada su similitud las dos muestras pertenecen a la misma clase. El k-NN también se basa en esa suposición, pero considera un mayor número de vecinos, lo cual permite una clasificación más robusta, menos sensible a los valores atípicos y al ruido de medición.

En el contexto del aprendizaje computacional, la clasificación por k-NN es un método de aprendizaje supervisado

de tipo "perezoso" que sólo requiere un mínimo esfuerzo en la etapa de entrenamiento y difiere todo el cómputo para la etapa de clasificación. Esto significa que inicialmente el k-NN no generaliza más allá de los datos de entrenamiento, y pospone esta acción hasta que el sistema debe clasificar datos nuevos.

Se considera que el conjunto de datos de entrenamiento contiene m muestras de n características del sistema. La i-ésima muestra es un vector de $\mathbb{R}^n$ de la forma $\mathbf{x}_i = (x_1, x_2, \ldots, x_n)^\mathsf{T}$ junto con una etiqueta que expresa su pertenencia a una de las q clases. Típicamente, las muestras $\mathbf{x}_i$ se agrupan en una matriz $\mathbf{X} \in \mathbb{R}^{m\times n}$; para las etiquetas de las clases se usan enteros positivos agrupados en un vector $\mathbf{y} \in \mathbb{R}^m$, donde y_i es la clase de la muestra $\mathbf{x}_i$. Se asume que las n características relevantes para la clasificación de las muestras han sido determinadas previamente mediante un proceso de selección (e.g. Dash and Liu, 1997) o extracción (e.g. Guyon et al., 2008). La **fase de entrenamiento** del algoritmo k-NN se limita a almacenar los vectores de características y las etiquetas de las clases. La Tabla 1 muestra la organización de los datos de entrenamiento requeridos.

Tabla 1. Datos de entrenamiento para el k-NN. Las etiquetas y_i son enteros positivos.

Muestra	Vector de características	Clase
1	$\mathbf{x}_1 = (x_{1,1}, x_{1,2}, \ldots, x_{1,n})^\mathsf{T}$	y_1
2	$\mathbf{x}_2 = (x_{2,1}, x_{2,2}, \ldots, x_{2,n})^\mathsf{T}$	y_2
$\vdots$	$\vdots$	$\vdots$
m	$\mathbf{x}_m = (x_{m,1}, x_{m,2}, \ldots, x_{m,n})^\mathsf{T}$	y_m

Posteriormente, en la **fase de clasificación** el algoritmo k-NN recibe nuevas muestras de clase desconocida y debe asignarles una de las q clases de los datos de entrenamiento, mediante algún proceso de inferencia. El algoritmo k-NN puede enmarcarse dentro de la teoría de decisión bayesiana, de modo que la clasificación de las nuevas observaciones se basa en hallar la clase con la mayor probabilidad a posteriori $P(c_j \,|\, \mathbf{x})$, $j = 1, 2, \ldots, q$, donde $\mathbf{x}$ es el vector de características de la muestra a clasificar y c_j representa la j-ésima clase. Mediante el teorema de Bayes, se demuestra (Martinez and Martinez, 2015) que la probabilidad a posteriori de cada clase está dada por

$$P(c_j \,|\, \mathbf{x}) = \frac{k_j}{k}, \tag{1}$$

donde k_j es el número de muestras del conjunto de entrenamiento que pertenecen a la j-ésima clase, considerando sólo las k muestras más cercanas a la muestra bajo prueba. Luego, la clase asignada por el clasificador se determina hallando la máxima probabilidad a posteriori:

$$\widehat{y}(\mathbf{x}) = \arg\max_j P(c_j \,|\, \mathbf{x}). \tag{2}$$

La salida $\widehat{y}(\mathbf{x})$ del algoritmo k-NN es una estimación de la clase verdadera $y(\mathbf{x})$, la cual, considerando (1), es el valor de y con la mayor frecuencia relativa entre los k vecinos más cercanos a $\mathbf{x}$. En la práctica, puede usarse sólo la frecuencia absoluta de las clases (el numerador en (1)) al maximizar la probabilidad a posteriori, dado que todas las probabilidades $P(c_j \,|\, \mathbf{x})$ tienen el mismo denominador. Para ejemplificar, en la Figura 1 se muestra una clasificación binaria (sólo dos clases) usando los 10 vecinos más cercanos (clasificación 10-NN). En este caso, 7 de los vecinos más cercanos al dato de prueba pertenecen a la Clase 2 y sólo 3 pertenecen a la Clase 1, de modo que $P(c_1 \,|\, \mathbf{x}) = 0.3$ y $P(c_2 \,|\, \mathbf{x}) = 0.7$. Por lo tanto, la salida del algoritmo k-NN sería $\widehat{y}(\mathbf{x}) = 2$, es decir, selecciona la Clase 2.

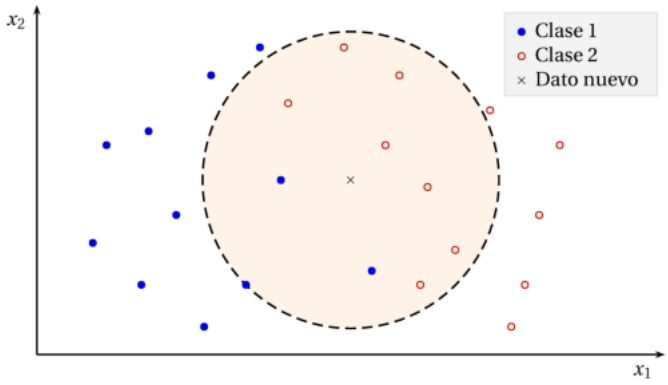

Figura 1. Clasificación 10-NN con distancia euclidiana, caso bidimensional.

Intuitivamente, para cuantificar la similitud entre dos muestras basta con medir su cercanía en el espacio de características, lo cual está determinado por la distancia entre los puntos que las representan. Bajo el enfoque euclidiano, la distancia entre dos muestras $\mathbf{x}$ y $\mathbf{x}'$ está determinada por

$$\begin{aligned} d(\mathbf{x}, \mathbf{x}') &= \sqrt{(x_1 - x_1')^2 + (x_2 - x_2')^2 + \cdots + (x_n - x_n')^2} \\ &= \sqrt{(\mathbf{x} - \mathbf{x}')^\mathsf{T}(\mathbf{x} - \mathbf{x}')}, \end{aligned} \tag{3}$$

y la región que encierra a las k muestras más cercanas al dato de prueba es una n-esfera (hiperesfera) en el espacio de características. En el caso bidimensional (véase la Figura 1), esta región es una circunferencia. Así en términos geométricos, el algoritmo k-NN consiste en calcular la densidad de puntos correspondiente a cada clase en la hiperesfera y seleccionar la clase con la mayor densidad. Sin embargo, aunque lo más frecuente para medir la similitud entre muestras es mediante los conceptos de distancia y norma usados en los espacios métricos (Zezula et al., 2006), la medida usada para cuantificar la similitud no tiene que ser forzosamente una distancia y, en caso de serlo, esta no tiene por qué ser euclidiana (Li et al., 2003; Cha, 2007). En los casos donde la similitud entre muestras no se mide mediante la distancia euclidiana, la región que encierra a los k vecinos más cercanos no tiene geometría esférica, y es más compleja que la mostrada en la Figura 1.

La distancia euclidiana (3) es un caso particular de la distancia de Minkowski,

$$d_p(\mathbf{x}, \mathbf{x}') = \sqrt[p]{\sum_{j=1}^{n} \left|x_j - x'_j\right|^p}, \tag{4}$$

con $p = 2$. Otras definiciones de distancia que particularizan (4) son la distancia Manhattan ($p = 1$) y la distancia de Chebychev ($p = \infty$). Cualquiera de estas métricas de distancia puede ser usada para darle sentido a la expresión "vecino más cercano".

Un aspecto a considerar cuando se usan la distancia euclidiana y otras distancias de Minkowski para medir la similitud entre muestras es que cada componente x_j del vector de características contribuye por igual al cuantificar la distancia. Esto puede dificultar la clasificación cuando existen algunas características x_j que son mucho más importantes que otras, o bien cuando se tienen muchas características irrelevantes. Por ejemplo, una característica relevante sería dominada por diez irrelevantes, de modo que resulta sensato asignar factores de ponderación a la contribución de $|x_j - x'_j|$ en el cálculo de la distancia. La necesidad de aplicar un escalamiento a cada término $|x_j - x'_j|$ en (4) es más evidente cuando las características x_j tienen diferentes escalas o unidades de medida. Una forma de escalamiento, que asigna las ponderaciones según la variabilidad de cada característica, es la distancia de Mahalanobis:

$$d_{\mathrm{M}}(\mathbf{x}, \mathbf{x}') = \sqrt{(\mathbf{x} - \mathbf{x}')^{\mathsf{T}} \mathbf{S}^{-1} (\mathbf{x} - \mathbf{x}')}, \tag{5}$$

donde $\mathbf{S}$ es la matriz de covarianza de los datos, de modo que $s_{i,j} = \mathrm{cov}(x_i, x_j)$. Sin la ponderación $\mathbf{S}^{-1}$, la distancia de Mahalanobis en (5) es la distancia euclidiana en (3).

Otra forma de cuantificar la similitud entre dos vectores, $\mathbf{x}_1$ y $\mathbf{x}_2$, es mediante el ángulo α entre ellos, o más precisamente del coseno del ángulo:

$$\cos\alpha = \frac{\mathbf{x}_1 \cdot \mathbf{x}_2}{\|\mathbf{x}_1\| \, \|\mathbf{x}_2\|}, \tag{6}$$

donde $\cdot$ expresa el producto interno (producto escalar) entre vectores y $\|\|$ representa la norma euclidiana. Esta medida tiene la peculiaridad de no depender de la magnitud de las muestras, solo de su orientación espacial. Como se mencionó en la introducción, respecto a la localización de fugas, los residuos de las presiones cuando las fugas ocurren en el mismo nodo de la red siguen aproximadamente la misma orientación espacial, independientemente de su magnitud, lo que sugiere usar el ángulo entre las muestras como medida de similitud. Para usarse como métrica de distancia, en el contexto del k-NN, la expresión en (6) debe ser cero cuando las muestras $\mathbf{x}_1$ y $\mathbf{x}_2$ sean iguales, de modo que se modifica de la siguiente manera:

$$d_{\cos}(\mathbf{x}_1, \mathbf{x}_2) = 1 - \frac{\mathbf{x}_1 \cdot \mathbf{x}_2}{\|\mathbf{x}_1\| \, \|\mathbf{x}_2\|} = 1 - \frac{\mathbf{x}_1^{\mathsf{T}} \mathbf{x}_2}{\sqrt{\mathbf{x}_1^{\mathsf{T}} \mathbf{x}_1} \sqrt{\mathbf{x}_2^{\mathsf{T}} \mathbf{x}_2}}. \tag{7}$$

La relación entre la distancia euclidiana y la distancia cosenoidal se analiza detalladamente en (Qian et al., 2004). Tanto (5) como (7) se usan como métricas de distancia para localización de fugas en el presente trabajo, pero se enfatiza la clasificación k-NN con distancia cosenoidal, pues esta demostró un desempeño superior respecto a la distancia euclidiana y a otras basadas en la distancia generalizada de Minkowski. El vector de características, $\mathbf{x}$, se construye con las presiones instantáneas en los nodos, o bien con las diferencias entre estas y las presiones nominales sin fuga.

Respecto a la elección de k (el número de vecinos a utilizar), esta se hace basándose en los datos disponibles, optimizando alguna medida del error de clasificación. Una forma de medir este error es mediante el **índice de pérdidas**:

$$\mathrm{loss} = 1 - \frac{m_{\mathrm{s}}}{m}, \tag{8}$$

donde m es el número total de muestras en los datos de prueba, y m_{s} es el número de estas muestras que resultan correctamente clasificadas con un valor dado de k. Para mejorar la robustez del clasificador, en la determinación del valor óptimo de k se usa validación cruzada iterativa (*K-fold cross validation*), de modo que el error de clasificación se calcula para un conjunto de muestras distinto al de entrenamiento. Por lo general, los valores pequeños de k (i.e. $k \to 1$) producen clasificadores muy sensibles al ruido de medición y a otras incertidumbres en el vector de características.

A diferencia de otras aplicaciones de clasificación donde solo es importante conocer qué fracción o porcentaje de las muestras el algoritmo identifica o estima correctamente, en la localización de fugas en redes también es importante cuantificar cada error de clasificación, pues si bien es grave que una fuga en el nodo 5 se localice erróneamente en el nodo 15, no lo es tanto si esta se localiza erróneamente en el nodo 6 (asumiendo que los nodos 5 y 6 sean adyacentes). Por ello, otra medida de error utilizada es el error topológico de localización, el cual indica a cuántos nodos de distancia se encuentra realmente la fuga respecto del nodo donde es detectada.

3. RESULTADOS

El algoritmo para localización de fugas mediante k-NN fue implementado en MATLAB® y probado con un banco de datos de fugas de la red de Hanoi, la que se muestra en la Figura 2 (Fujiwara and Khang, 1990). Esta red de prueba consta de 32 nodos (31 uniones y un depósito) y 34 líneas de tubería que suman una longitud de 39 420 metros. El banco de datos fue generado mediante una simulación con el software EPANET (Rossman, 2000) usando un modelo hidráulico que considera la presión disponible en el depósito, la geometría de las tuberías y su rugosidad, así como las demandas en los nodos de consumo. Para simular las fugas, mediante la interfaz del *EPANET/MATLAB Toolkit* (Eliades et al., 2016) se manipularon las demandas en los nodos de consumo, incrementando la demanda base por una cantidad igual al caudal de la fuga simulada. En cada nodo de la red se simularon fugas de distintas magnitudes, considerando caudales $Q_{\mathrm{fuga}} = \{1, 2, \ldots, 40\}$ l/s. En esta forma se construyó una matriz de presiones nodales con 1 240 escenarios hipotéticos de fuga, que corresponden a las 40 diferentes magnitudes de fuga por cada uno de los 31 nodos de unión.

Este conjunto de datos se particionó en 4 subconjuntos (uno para entrenamiento y tres para prueba), con 10 muestras cada uno; luego, se usó validación cruzada de 4 iteraciones (4-*folding*) para determinar el valor óptimo de k y medir el desempeño del clasificador calculando los índices de pérdidas.

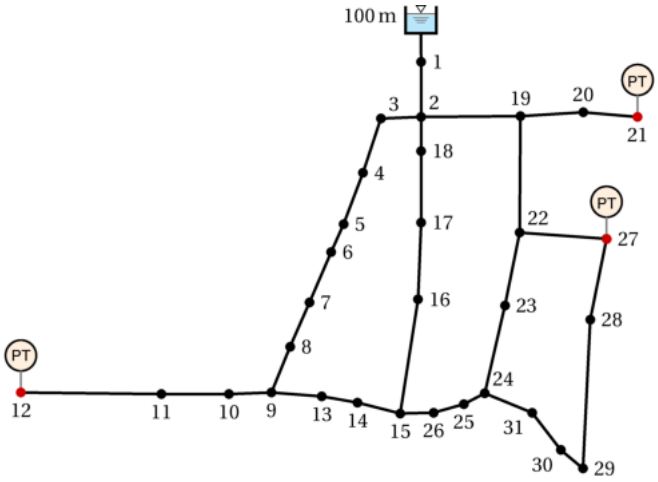

Figura 2. Red de Hanoi. Los nodos marcados con PT contienen sensores de presión.

Para asemejarse a una aplicación real, donde solo se monitorea un número limitado de nodos de la red, en las pruebas realizadas con el clasificador k-NN se considera que únicamente están disponibles las mediciones de presión en los nodos 12, 21 y 27. De esta manera, el vector de características referido en la sección previa (la entrada del clasificador) quedó definido por

$$\mathbf{x} \stackrel{\text{def}}{=} \mathbf{P} = (P_{12}, P_{21}, P_{27})^{\mathsf{T}}. \tag{9}$$

También se hicieron pruebas con un vector de características construido con los residuos de las presiones,

$$\mathbf{x} \stackrel{\text{def}}{=} \mathbf{r} = (r_{12}, r_{21}, r_{27})^{\mathsf{T}}, \tag{10}$$

donde $r_j = P_j - \overline{P}_j$ es el cambio negativo de presión en el j-ésimo nodo debido a la fuga, respecto a la presión nominal (sin fuga) $\overline{P}_j$. Las presiones nominales también fueron calculadas mediante simulación con el modelo hidráulico de la red. En los resultados presentados a continuación se especifica cuánto cambia el desempeño del localizador de fugas al usar $\mathbf{P}$ o $\mathbf{r}$. Para estudiar la robustez del sistema, además del conjunto de prueba (matriz de presiones nodales) referido al inicio de esta sección se generó otro conjunto de prueba adicionando ruido gaussiano a las presiones obtenidas por simulación. La razón señal/ruido (SNR) en las presiones de este segundo conjunto de prueba se estableció en 80 dB. También se construyó un tercer conjunto de prueba más ruidoso con SNR = 60 dB. Se calcularon dos índices de pérdidas para cuantificar el desempeño del clasificador k-NN, el **índice de pérdidas por resustitución** (r-loss) y el **índice de pérdidas con validación cruzada** (cv-loss); el primero mide la incapacidad del clasificador k-NN para localizar correctamente las fugas conocidas (conjunto de entrenamiento), mientras que el segundo mide la incapacidad para localizar fugas desconocidas (conjuntos de prueba). También se calculó la **distancia topológica media** (ATD), la cual representa el número promedio de nodos que separan la posición estimada de la fuga de su posición real.

La Tabla 2 resume los resultados de la localización de fugas en la red de Hanoi al usar k-NN con métrica cosenoidal. Se reporta el desempeño tanto con mediciones sin ruido como ruidosas, considerando dos diferentes vectores de características: las presiones en los nodos ($\mathbf{P}$) y sus residuos ($\mathbf{r}$). En el k-NN con métrica cosenoidal, el desempeño del clasificador que usa residuos supera, por mucho, al que utiliza directamente las presiones.

Tabla 2. Desempeño en la localización de fugas mediante k-NN con distancia cosenoidal.

Error	Sin ruido		SNR = 80 dB		SNR = 60 dB	
	Usa **P**	Usa **r**	Usa **P**	Usa **r**	Usa **P**	Usa **r**
r-loss	0.4863	0.0016	0.4468	0.0323	0.4895	0.2694
cv-loss	0.6008	0.0024	0.6081	0.0452	0.6879	0.3153
ATD	1.3274	1.0645	1.3548	1.1355	1.7097	1.5177

En las pruebas se encontró que el valor óptimo de k (número de vecinos cercanos) para obtener el mejor desempeño en la localización de fugas depende del nivel de ruido en las mediciones. Considerando el índice de pérdidas con validación cruzada como criterio de optimalidad, y con la métrica de distancia cosenoidal, el mejor desempeño en ausencia de ruido se obtuvo con $k = 2$; para condiciones ruidosas con SNR = 80 dB, el valor óptimo resultó $k = 4$, mientras que para SNR = 60 dB el mejor resultado corresponde a $k = 7$. En la Figura 3 se muestra la variación del índice de pérdidas respecto de k, para el caso de mediciones ruidosas con SNR = 80 dB.

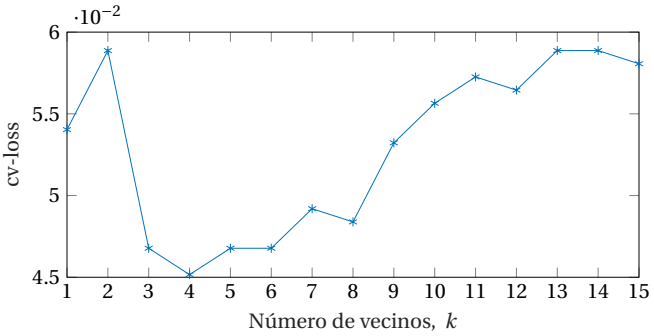

Figura 3. Índice de pérdidas por validación cruzada para diferente número de vecinos, considerando SNR = 80 dB.

El desempeño del k-NN con métrica cosenoidal, resultó bastante superior al obtenido con las distancias euclidiana y de Mahalanobis. Por ejemplo, para el k-NN con distancia euclidiana el menor cv-loss fue de 0.5202 sin ruido y de 0.5266 cuando se considera ruido con SNR = 80 dB; para el k-NN con distancia de Mahalanobis el menor cv-loss fue de 0.4831 sin ruido y de 0.4879 cuando se considera ruido

con SNR = 80 dB. En ambos casos el mejor desempeño se obtuvo con solo dos vecinos ($k = 2$) y no existe diferencia significativa si se utilizan directamente las presiones nodales en lugar de sus residuos.

Comparado con el desempeño en la localización de fugas del método basado en la matriz de sensibilidad que considera sólo la máxima correlación, el clasificador k-NN con métrica cosenoidal mostró un desempeño superior en condiciones altamente ruidosas, pero no en condiciones ideales o con poco ruido. Las condiciones de ruido para las que el k-NN cosenoidal mostró un desempeño superior son SNR < 48 dB, usando solo las tres mediciones de presión indicadas en la Figura 2.

4. CONCLUSIONES

El clasificador k-NN con métrica cosenoidal demostró un buen desempeño en la localización de fugas, considerando su exactitud y su robustez. Con esta medida de distancia se obtuvieron mejores resultados que con las métricas euclidiana y de Mahalanobis. La localización de fugas mediante k-NN con métrica cosenoidal también mostró un mejor desempeño que el método de la máxima correlación, en condiciones altamente ruidosas. En trabajos futuros se espera probar esta metodología usando mediciones físicas en redes de distribución de agua con un mayor número de nodos. Además, considerando que el planteamiento actual requiere de un modelo hidráulico de la red bien calibrado para obtener los datos de entrenamiento, a futuro se pretende desarrollar metodologías menos dependientes del modelo, en la búsqueda de técnicas de localización de fugas completamente basadas en datos.

REFERENCIAS

Blesa, J., Nejjari, F., and Sarrate, R. (2016). Robust sensor placement for leak location: analysis and design. *Journal of Hydroinformatics*, 18(1), 136–148.

Casillas, M., Garza-Castañón, L., and Puig, V. (2015). Optimal sensor placement for leak location in water distribution networks using evolutionary algorithms. *Water*, 7(11), 6496–6515.

Casillas, M.V., Garza-Castañón, L.E., and Puig, V. (2013). Extended-horizon analysis of pressure sensitivities for leak detection in water distribution networks: Application to the barcelona network. In *2013 European Control Conference (ECC)*, 401–409. IEEE.

Cha, S.H. (2007). Comprehensive survey on distance/similarity measures between probability density functions. *International Journal of Mathematical models and Methods in Applied Sciences*, 1(4), 300–307.

Dash, M. and Liu, H. (1997). Feature selection for classification. *Intelligent Data Analysis*, 1(1), 131–156. doi:10.1016/S1088-467X(97)00008-5.

Eliades, D.G., Kyriakou, M., Vrachimis, S., and Polycarpou, M.M. (2016). EPANET-MATLAB Toolkit: An Open-Source Software for Interfacing EPANET with MATLAB. In *Proc. 14th International Conference on Computing and Control for the Water Industry (CCWI)*, 8. The Netherlands. doi:10.5281/zenodo.831493.

Ferrandez-Gamot, L., Busson, P., Blesa, J., Tornil-Sin, S., Puig, V., Duviella, E., and Soldevila, A. (2015). Leak localization in water distribution networks using pressure residuals and classifiers. *IFAC-PapersOnLine*, 48(21), 220–225.

Fujiwara, O. and Khang, D.B. (1990). A two-phase decomposition method for optimal design of looped water distribution networks. *Water resources research*, 26(4), 539–549.

Guyon, I., Gunn, S., Nikravesh, M., and Zadeh, L.A. (2008). *Feature extraction: foundations and applications*, volume 207. Springer.

Li, M., Chen, X., Li, X., Ma, B., and Vitányi, P. (2003). The similarity metric. In *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, 863–872. Society for Industrial and Applied Mathematics.

Martinez, W.L. and Martinez, A.R. (2015). *Computational Statistics Handbook with MATLAB*. Chapman & Hall/CRC Computer Science & Data Analysis. Chapman and Hall/CRC, 3 edition.

OECD (2016). Water Governance in Cities. *OECD Studies on Water*.

Perez, R., Sanz, G., Puig, V., Quevedo, J., Cuguero Escofet, M.A., Nejjari, F., Meseguer, J., Cembrano, G., Mirats Tur, J.M., and Sarrate, R. (2014). Leak localization in water networks: A model-based methodology using pressure sensors applied to a real network in barcelona [applications of control]. *IEEE Control Systems Magazine*, 34(4), 24–36. doi:10.1109/MCS.2014.2320336.

Puig, V., Ocampo-Martínez, C., Pérez, R., Cembrano, G., Quevedo, J., and Escobet, T. (eds.) (2017). *Real-time Monitoring and Operational Control of Drinking-Water Systems*. Springer International Publishing. doi:10.1007/978-3-319-50751-4.

Pérez, R., Puig, V., Pascual, J., Quevedo, J., Landeros, E., and Peralta, A. (2011). Methodology for leakage isolation using pressure sensitivity analysis in water distribution networks. *Control Engineering Practice*, 19(10), 1157 – 1167. doi:10.1016/j.conengprac.2011.06.004.

Qian, G., Sural, S., Gu, Y., and Pramanik, S. (2004). Similarity between euclidean and cosine angle distance for nearest neighbor queries. In *Proceedings of the 2004 ACM symposium on Applied computing*, 1232–1237. ACM.

Rossman, L.A. (2000). EPANET 2: Users manual. Technical Report EPA/600/R-00/057, US Environmental Protection Agency.

Soldevila, A., Blesa, J., Tornil-Sin, S., Duviella, E., Fernandez-Canti, R.M., and Puig, V. (2016). Leak localization in water distribution networks using a mixed model-based/data-driven approach. *Control Engineering Practice*, 55, 162 – 173. doi:10.1016/j.conengprac.2016.07.006.

Zezula, P., Amato, G., Dohnal, V., and Batko, M. (2006). *Similarity Search: The Metric Space Approach*, volume 32 of *Advances in Database Systems*. Springer Science & Business Media. doi:10.1007/0-387-29151-2.

