Advances in Engineering and Innovation

Advances in Engineering and Innovation
Vol. 8, No. 17, pp. 98-106
Enero - Junio 2023

www.progreso.tecnm.mx/revistaAEI
ISSN: 2448-685X

Localización de fugas en redes de distribución por medio de un clasificador de bosque aleatorio

Leak localization in water distribution networks using random forest classifier

Anuar Badillo Olvera[1*], Julio Zenón García Cortes[1], Larbi Djilali[2], Erick del Jesus Tamayo Loeza[3] y Ramón Salvador Mezquita Martínez[3]

[1]*Tecnológico Nacional de México campus Zacatecas Norte, carretera a González Ortega Km 3, C.P. 98400, Ap. postal 178, Río Grande, Zacatecas, México.*

[2]*Universidad Autónoma del Carmen, Av. 56 No. 4, Esq. Avenida Concordia, colonia Benito Juárez C.P. 24180, Ciudad del Carmen, Campeche, México.*

[3]*Tecnológico Nacional de México campus Progreso, Boulevard Tecnológico de Progreso por 62 SN, C.P. 97320, Progreso, Yucatán, México.*

[*]*Corresponding author: anuar.bo@zacatecasnte.tecnm.mx*

***Resumen*—Este artículo presenta un algoritmo de clasificación basado en el algoritmo de bosque aleatorio que tiene por objetivo el determinar la ubicación aproximada de una fuga en una red de distribución de agua. En esta técnica, residuos obtenidos de la comparación entre las medidas de presión en los nodos de la red y el comportamiento nominal estimado a partir del modelo matemático. Dichos se analizan por medio de un algoritmo de clasificación con el objetivo de asociar la información residual a una ubicación aproximada de fuga.. El desempeño del algoritmo es evaluado y comparado con otros algoritmos de clasificación en un problema de tipo que comprende ciento veintiséis nodos, ocho válvulas, dos tanques, dos depósitos, tres bombas y ciento sesenta y ocho tuberías, sujeto a cuatro patrones de demanda variable. los resultados ilustran la mejora de la precisión del algoritmo propuesto sobre las otras técnicas (Algoritmo de vecinos cercanos y máquinas de vector soporte).**

***Palabras clave*—Fuga, Red de distribución, Localización de fugas, Reducción de pérdidas, Presiones de red.**

***Abstract*—This paper presents a Random Forest classifier technique for determining the approximate location of a leak in a Water Distribution Network (WDN). In this technique residuals obtained from the comparison between the pressure measurements and the nominal behavior estimated from the mathematical model are analyzed as a classification task in order to associate the residual information**

Recibido: 19/Octubre/2022
Aceptado: 09/Diciembre/2022

to an approximate leak location. The algorithm performance is evaluated and compared with other classification algorithms on a benchmark problem comprising by one hundred and twenty-six nodes, eight valves, two tanks, two reservoirs, three pumps, and one hundred and sixty-eight pipes, subject to four variable demand patterns. the results illustrate the improved accuracy of the proposed algorithm over the other techniques (Near Neighbors Algorithm and Support Vector Machines).

***Keywords*—Leakage, Water distribution networks, Fault localization, Loss reduction, Pressure-Driven Deman.**

I. Introducción

Las pérdidas de agua es un problema al que se enfrentan los organismos de administración del agua de todo el mundo, debido a las posibles consecuencias relacionadas con los daños a la seguridad, la economía y el medio ambiente. Además, las fugas de agua pueden ocurrir debido al deterioro debido al tiempo de la infraestructura. En consecuencia, porcentajes de fugas (pérdidas) comprendidos en rangos entre el 30 y el 50 % son comunes en los Sistemas de Distribución de Agua (SDA) [World Water Assessment Programme United Nations and UN-Water, 2009], [Puig et al., 2016]. En ese sentido, se han propuesto métodos de detección y localización de fugas basados en diferentes enfoques para abordar este problema. Uno de estos enfoques es el basado en modelos, donde se utilizan mediciones de flujo y presión en los nodos de la red, modelos hidráulicos y diferentes métodos de estimación para realizar línea la tarea de detección y localización de fugas. Estos métodos formulan el problema de localización de fugas en función de diferentes consideraciones de diseño y supuestos matemáticos, como ejemplo, en [Pudar and Liggett, 1992] se aborda el problema de fugas como un problema de estimación paramétrica por medio de la técnica de mínimos cuadrados. Sin embargo, estos métodos no toman en consideración las incertidumbres en la estimación de la demanda y las perturbaciones en las mediciones, las cuales se presentan en aplicaciones reales. Otros trabajos trata el problema de localización de fugas como una tarea de clasificación [Valizadeh et al., 2009], [Leu and Bui, 2016]. Estos enfoques pueden considerarse basados datos, ya que solo es necesario utilizar datos experimentales sin necesidad de modelos matemáticos. Sin embargo, el rendimiento de estos métodos está fuertemente relacionado con el proceso de entrenamiento, pero en aplicaciones reales, puede ser difícil obtener conjuntos de datos que cubran todas las fallas posibles. Considerando esto, los autores en [Puig et al., 2016], [Sarrate et al., 2014] proponen un método mixto basado en modelos y controlado por datos, donde se considera un residuo el cual se genera a partir del estado nominal y la condición de falla, y se analiza utilizando técnicas de clasificación. En este enfoque, no es necesario que los conjuntos de datos cubran todas las posibles fallas como en los métodos basados en datos puros, ya que el esquema basado en modelos permite generar las posibles fallas. Buscando mejorar el rendimiento general del modelo mixto, este artículo propone el diseño de un clasificador basado en la técnica de arboles de decisión (RF, del inglés: Random Forest) con datos de un caso de estudio.

El modelo hidráulico del caso de estudio corresponde a la *Red 1* utilizada para evaluar el desempeño del algoritmo en el concurso *"La Batalla de las Redes de Sensores de Agua"*, BWSN [Ostfeld et al., 2008]. El desempeño de el esquema de localización de fugas propuesto se evalúa en un escenario perturbado por ruido e incertidumbres en la estimación de la demanda.

El resto del documento está organizado de la siguiente manera: la sección II presenta los preliminares compuestos por: el modelo de distribución de agua, describe el número de sensor considerado y el procedimiento de colocación que se utiliza para definir el esquema de instrumentación relacionado con la fuga propuesta algoritmo de localización. El esquema de clasificación propuesto para abordar la localización de fugas para se presenta en la Sección III. La sección IV, presenta el algoritmo de clasificación propuesto a través de los resultados obtenidos en un problema de referencia. Finalmente, se exponen las respectivas conclusiones.

II. Preliminares

Esta sección presenta el esquema de simulación hidráulica, el modelo de fuga, la selección del número y posición de los sensores utilizados en este trabajo.

II-A. Ecuaciones gobernantes

El análisis de caudal y presión de un red de distribución se puede realizar desde diferentes enfoques de modelo, como: modelos inerciales, transitorios, estáticos y cuasiestáticos. Los modelos inercial y transitorio se caracterizan por ecuaciones diferenciales parciales, mientras que el modelo estático analiza estados estacionarios a través de un sistema de ecuaciones algebraicas. Los modelos estáticos se pueden extender a diferentes períodos de simulación con la superposición de simulaciones estáticas en el tiempo (modelos cuasi-estáticos) con diferentes condiciones de frontera. En general, para aplicaciones prácticas se utilizan los modelos estático y cuasi-estático, ya que con este tipo de modelos se puede abordar la gestión diaria de la red (análisis, diseño y operación). Además, para Redes grandes, los efectos transitorios se disipan rápidamente y los modelos transitorios son computacionalmente costosos [Cabrera and Vela, 2013].

El análisis de red estática se basa en las ecuaciones de masa (1) y energía (2):

$$\sum_{j=1}^{P} Q_{ij} - q_i = 0 \quad \text{for } i = 1, \ldots, N \tag{1}$$

donde j es el conjunto de tuberías conectadas al nodo i, P es el número total de tuberías, q_i es la demanda en el nodo i, Q_{ij} es el gasto en el nodo i de la tubería j, y N es el número total de nodos. Por otro lado, la ecuación de la energía involucra balances de energía, trabajo, potencia y máquinas que interactúan con los fluidos que fluyen y la red de distribución de agua. La ecuación de energía, del nodo i en la tubería j, se puede expresar como:

$$H_i - H_j = h_{ij} = rQ_{ij}^n + mQ_{ij}^2 \tag{2}$$

donde n es un exponente de flujo que depende de la ecuación de fricción, r es el coeficiente de fricción, m es la pérdida local, H_i y H_j es la cabeza de presión en el nodo i-ésimo y la cabeza al final del tubo j-th, respectivamente [Rossman, 2001]. La cabeza de presión agregada por las bombas se describe mediante una ley de potencia:

$$h_p = -\omega^2(h_0 - r_p(Q_{ij}/\omega)^{\eta_p}) \tag{3}$$

donde ω es un parámetro asociado a la de velocidad relativa, r_p y η_p son los coeficientes de la curva de la bomba, y h_0 es la cabeza de cierre de la bomba. El uso de (1-3) para describir el comportamiento de un red conduce a un sistema de ecuaciones no lineales que puede aproximarse mediante un método numérico [Rossman, 2001].

El modelado tradicional de redes de agua utiliza como supuesto que las presiones en la red son función de la demanda constante. Este enfoque de conduce a buenos resultados cuando las presiones en la red son suficientes. Sin embargo, en escenarios donde se presentan condiciones de baja presión por ejemplo en escenarios de falla donde los consumidores no siempre reciben su demanda y, en ocasiones, generan presiones negativas en los resultados de la simulación, siendo una consideración poco realista. De esta forma, trabajos como en [Giustolisi et al., 2008], [Klise et al., 2017], [Todini, 2010] proponen una simulación de Demanda en función de la Presión (PDD) para obtener una representación más realista de las redes de distribución bajo escenarios de falla. Considerando un escenario de fugas de una red que comprende n_p tuberías con descargas desconocidas, n_n nodos con alturas desconocidas (nodos internos) y n_t nodos con alturas conocidas, por ejemplo, niveles de tanque, la simulación PDD se puede describir mediante el siguiente sistema de ecuaciones basadas en (1) y (2):

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{Q} \\ \mathbf{H} \end{bmatrix} = \begin{bmatrix} -\mathbf{A}_{10}\mathbf{H}_0 \\ -q \end{bmatrix}, \tag{4}$$

donde $\mathbf{H}_0 = [H_{0,1}, H_{0,2}, \ldots, H_{0,n_t}]^T$ es un vector columna de cabezas nodales conocidas, $\mathbf{Q} = [Q_1, Q_2, \ldots, Q_{n_p}]^T$ y $\mathbf{H} = [H_1, H_2, \ldots, H_{n_n}]^T$ es una columna vector de caudales y alturas de tubería desconocidos, respectivamente. En (4), $\mathbf{A}_{11}$ es una matriz diagonal de orden n_p cuyos elementos se definen por las pérdidas de carga como:

$$A_{11}(k,k) = nr|Q_{ij}|^{n-1} + 2m|Q_{ij}|, \tag{5}$$

para tuberías y:

$$A_{11}(k,k) = -\omega(h_0 - r_p(Q_{ij}/\omega^{n_p}))/Q_{ij}, \tag{6}$$

para bombas, para $k \in 1, n_p$; $k \in 1, n_t$; $j \in 1, n_t$. Por otro lado, la topología de la red se describe mediante la sub matriz de incidencia $\overline{\mathbf{A}_{12}}$, definida de la siguiente manera:

$$\overline{\mathbf{A}_{12}}(i,j) = \begin{cases} -1 & \text{si el flujo de la tubería } j \text{ sale del nodo } i, \\ 0 & \text{si la tubería } j \text{ no está conectada al nodo } i, \\ +1 & \text{si el flujo de la tubería } j \text{ ingresa al nodo } i. \end{cases} \tag{7}$$

La submatriz de incidencia general se puede dividir en dos submatrices: $\overline{\mathbf{A}}_{12} = [\mathbf{A}_{12} \vdots \mathbf{A}_{10}]$ relacionando las tuberías con los nodos con cabeza de presión desconocida ($\mathbf{A}_{12}$).

II-B. Número y ubicación del sensor

El desempeño de un sistema de monitoreo depende en gran medida de las mediciones disponibles. Por ello, antes de implementar un método de localización de fugas, la selección del número de sensores y su posición es una tarea fundamental. El uso de un sensor en cada nodo de la red podría ser un escenario ideal para el desempeño del monitoreo. Sin embargo, este escenario puede estar involucrando un gran costo de instrumentación y en general las agencias gestoras del agua buscan una configuración eficiente donde se maximice el monitoreo con un número reducido de sensores, evitando un gran costo de instrumentación [Sarrate et al., 2014].

La colocación de M sensores en N nodos potenciales con $M < N$ requiere un tiempo exponencial para un método exhaustivo, incluso para un método de tiempo polinomial resulta inviable para un WDN de tamaño intermedio [Gamboa-Medina and Reis, 2017]. De esta manera, [Sarrate et al., 2014] propone dos pasos para abordar el problema del número de sensores y su ubicación:

- Primero, para evitar información redundante y reducir los sensores candidatos iniciales, se utiliza una técnica de agrupamiento para determinar la similitud entre las firmas de falla (la diferencia entre la presión nominal y las condiciones de falla). Una vez dividido el sensor en grupos naturales, el

conjunto del número máximo de sensores instalados será igual al número de grupos.
- Dando el número máximo de sensores instalados, el problema de ubicación se aborda como un problema de optimización, buscando las mejores ubicaciones posibles de sensores que maximizan el rendimiento de monitoreo calculado a través de una función de costo.

En este trabajo, los datos (objetos) analizados están dados por una matriz de sensibilidad como en [Pudar and Liggett, 1992], [Puig et al., 2016], [Casillas et al., 2015] que en un esquema de simulación considera todos los posibles sensores instalados en el sistema, por lo que esta matriz contiene las posibles desviaciones de presión en un WDS con fugas presentes. Esta matriz está compuesta por vectores de sensibilidad o síntoma $\mathbf{s}_j$ que contienen una diferencia normalizada entre la presión de cabeza de una simulación sin fugas $\hat{H}_i$ y la presión de cabeza $\hat{H}_i^{f_j}$ que representa la presión de cabeza en cada nodo de la red bajo los efectos de la magnitud de la fuga f_j. La matriz de sensibilidad se puede expresar como:

$$\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N] \text{ con } \mathbf{s}_j = \begin{bmatrix} \frac{\hat{H}_1{}^{f_j} - \hat{H}_1}{f_j} \\ \vdots \\ \frac{\hat{H}_N^{f_j} - \hat{H}_N}{f_j} \end{bmatrix}. \quad (8)$$

Para determinar el número y la posición del sensor, la información de la matriz de sensibilidad se analiza en dos pasos. En el primer paso se determina el número de sensor a través de una técnica de agrupamiento y en el segundo paso se aborda el problema de ubicación del sensor.

En el uso de una técnica de agrupamiento para determinar el número de sensores, es necesario considerar las dificultades que se presentan en un escenario realista, como variaciones temporales de la demanda, ruido en las mediciones y fugas de diferente magnitud. Bajo este escenario, es difícil definir una similitud entre objetos, ya que pueden estar simultáneamente cerca de diferentes grupos. Una opción para abordar este tipo de problemas son los métodos *Fuzzy C-Means* donde los objetos pueden pertenecer a más de un clúster con cierto grado de pertenencia. Asimismo, es importante señalar que en las técnicas de agrupamiento, de forma no supervisada, no existe una información previa sobre el número de agrupamientos naturales. Teniendo esto en cuenta, la mayoría de los métodos calculan un índice de validez de varias particiones para seleccionar el número de clúster apropiado, [Masson and Denœux, 2008]. En este trabajo, el problema de agrupamiento se resuelve utilizando tres técnicas basadas en agrupamiento difuso: el clásico Fuzzy C-means (FCM) [A Miyamoto et al., 2008], el Evidential C-Means (ECM) [Masson and Denœux, 2008] y el Kernelized Fuzzy C-means (FCM) [A Miyamoto et al., 2008] Algoritmos de medios (KFCM) [Girolami, 2002]. Con el fin de obtener una decisión más informada, en este trabajo se comparan estas técnicas mediante el índice de validez propuesto en [A Miyamoto et al., 2008], [Masson and Denœux, 2008] y [Girolami, 2002] para los métodos FCM, ECM y KFCM, respectivamente.

Como se mencionó anteriormente, una vez que el número de sensor se define a través de la técnica de agrupación, es necesario abordar el problema de ubicación del sensor. Este artículo implementa el método *basado en proyecciones* presentado en [Casillas et al., 2015] para seleccionar la posición del sensor. Este método considera la matriz de sensibilidad $\mathbf{S}$ descrita anteriormente y la matriz residual $\mathbf{R}$ [Pudar and Liggett, 1992], [Puig et al., 2016]. La matriz residual $\mathbf{R}$ se construye de la misma manera que $\mathbf{S}$, pero con una magnitud de fuga diferente f_j y sin normalización. Después, la configuración del candidato del sensor se mapea en un vector binario $\mathbf{p} = [p_1, p_2, \ldots, p_N]$, donde:

$$p_i = \begin{cases} p_i = 0 & \text{Si la presión es medida en el nodo,} \\ p_i = 1 & \text{de otro modo.} \end{cases} \quad (9)$$

A partir del vector binario se construye una matriz diagonal $\mathbf{P} = \text{diag}(p_1, p_2, \ldots, p_N)$ y se calcula una medida de aislabilidad basada en proyecciones vectoriales normalizadas [Casillas et al., 2015]:

$$\psi_{k,j}(\mathbf{p}) = \frac{S_{2,k}^T P(\mathbf{p}) S_{1,j}}{\|P(\mathbf{p}) S_{2,k}\| \, \|P(\mathbf{p}) S_{1,j}\|}, \quad (10)$$

A partir de las proyecciones vectoriales es posible construir la matriz (Ψ), como:

$$\Psi(\mathbf{p}) = \begin{bmatrix} \psi_{11}(\mathbf{p}) & \cdots & \psi_{1N}(\mathbf{p}) \\ \vdots & \ddots & \vdots \\ \psi_{N1}(\mathbf{p}) & \cdots & \psi_{NN}(\mathbf{p}) \end{bmatrix}. \quad (11)$$

Para evaluar la calidad del índice de error de configuración candidato se introduce como:

$$\varepsilon_i(\mathbf{p}) = \begin{cases} 0 & \text{if} \quad \psi_{ii}(\mathbf{p}) = \max(\psi_{i1}(\mathbf{p}), \ldots, \psi_{iN}(\mathbf{p})), \\ 1 & \text{de otro modo,} \end{cases} \quad (12)$$

Después, un índice global que tiene en cuenta todos los nodos se calcula como:

$$\overline{\varepsilon}(\mathbf{p}) = \sum_{i=1}^{m} \frac{\varepsilon_i(\mathbf{p})}{m}, \quad (13)$$

Finalmente, la tarea de colocación de sensores se formula como un problema de optimización de la siguiente manera:

$$\begin{aligned} \min_{p} \quad & \overline{\varepsilon}(\mathbf{p}) \\ \text{s.t.} \quad & \sum_{i=1}^{m} \mathbf{p}_i = s_n, \end{aligned} \quad (14)$$

donde p es el vector de la posición del sensor candidato y s_n es el número de sensor a colocar. Para abordar el problema de optimización (14), se implementa el algoritmo genético (GA) como en [Casillas et al., 2015].

II-C. Diagnóstico de fugas en redes de distribución

Al igual que en líneas de conducción, la localización de fugas en redes,abordada desde un enfoque por modelo, radica en la comparación de mediciones de presión distribuidas en puntos estratégicos de la tubería con un modelo matemático que describe el comportamiento teórico de la red. Es por ello que el punto de partida para la implementación de un sistema de localización de fugas en redes, parte del posicionamiento estratégico y la selección del número de sensores necesarios para una correcta identificación. En general el posicionamiento y la selección del número de sensores es planteado como problema de optimización.

El posicionamiento de M sensores en N nodos potenciales requiere tiempo exponencial para un método exhaustivo, incluso para un método polinomial resulta inviable para una red de tamaño medio [Gamboa-Medina and Reis, 2017]. Considerando esto, en este artículo se aborda este problema como [Sarrate et al., 2014]. Bajo el esquema de instrumentación obtenido, es abordado el problema de localización de fugas considerando que ocurre una fuga a la vez y que sólo ocurre en los nodos, ya que en el modelado las demandas son asignadas a los nodos. Bajo estas consideraciones, la localización de fugas en redes se basa en la comparación de las estimaciones de presión en los nodos $\hat{H}_i$ estimadas mediante el modelo matemático HDD, definido en anteriormente con las mediciones de presión H obtenidas de los sensores distribuidos en la red, de ésta comparación es generado un residuo (r) el cual es analizado con el objetivo de extraer síntomas y localizar la fuga. Considerando la complejidad de la red, para abordar el problema la localización de la fuga, en [Sarrate et al., 2014], [Soldevila et al., 2017], se propone el uso de técnicas de reconocimiento de patrones, donde el problema es formulado como un problema supervisado de clasificación multivariable como:

Dado un conjunto $X = \{(r_1, l_1), (r_2, l_2), \ldots, (r_N, l_N)\}$ de datos residuales donde $\mathbf{r}_i \in \mathbb{R}^{sn}$, sn es el número de sensores, $\mathbf{r}_i$ es la presión residual y $l_i \in \mathbb{L}$ es la clase que indica el nodo donde la fuga ha ocurrido, el problema de localización o aislamiento de fugas consiste en encontrar una función:

$$g : \mathbb{R}^{sn} \leftarrow \{1, 2, \ldots, N\}, \tag{15}$$

es decir, encontrar una función de clasificación que para cualquier objeto $x \in \mathbb{R}sn$ a ser clasificado, obtenga como salida la etiqueta de clase li asociada a la posición de una fuga. Aunado a esto, en un escenario práctico es importante considerar que las mediciones son perturbadas con ruido, existe incertidumbre en los patrones de demanda y las fugas pueden aparecer en cualquier nodo con diferente magnitud. Teniendo en cuenta todos estos efectos, el clasificador diseñado debe poder localizar la fuga real presente en la red y a su vez debe ser robusto a al ruido de las mediciones. El esquema de localización de la fuga puede ser resumido mediante el esquema de la figura 1.

II-D. Clasificador de Bosque Aleatorio

Un árbol de decisión es un clasificador expresado por medio de particiones recursivas del espacio de instancias. El árbol de decisión se conforma por nodos que unen un conjunto de posibles decisiones basadas en construcciones lógicas, donde cada nodo interno divide el espacio de la instancia en dos o más subespacios de acuerdo con una determinada función acorde con los valores de los atributos de entrada. En el caso más simple y frecuente, cada prueba considera un solo atributo, de modo que el espacio de la instancia se divide de acuerdo con el valor de los atributos. En Figura 2 se presenta un ejemplo simple de un árbol de decisión.

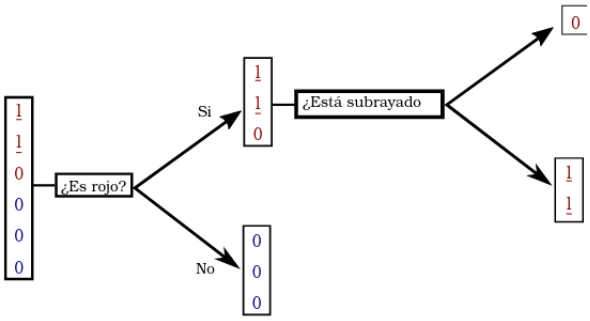

Figura 2. Ejemplo simple de un árbol de decisión. Fuente: Elaboración propia.

En la técnica de clasificación por medio de bosque aleatorio (RF, del inglés Random Forest) emplea un conjunto de árboles de decisión, donde cada clasificador es generado usando un vector de muestras aleatorio, independiente al vector de entrada pero de la misma dimensión. De cada árbol se obtiene un resultado unitario que indica la pertenencia de la clase de entrada a un determinado grupo. La clase final se determina mediante el promedio de los clasificadores [Pal, 2005]. La Figura 3 muestra un ejemplo de un bosque aleatorio.

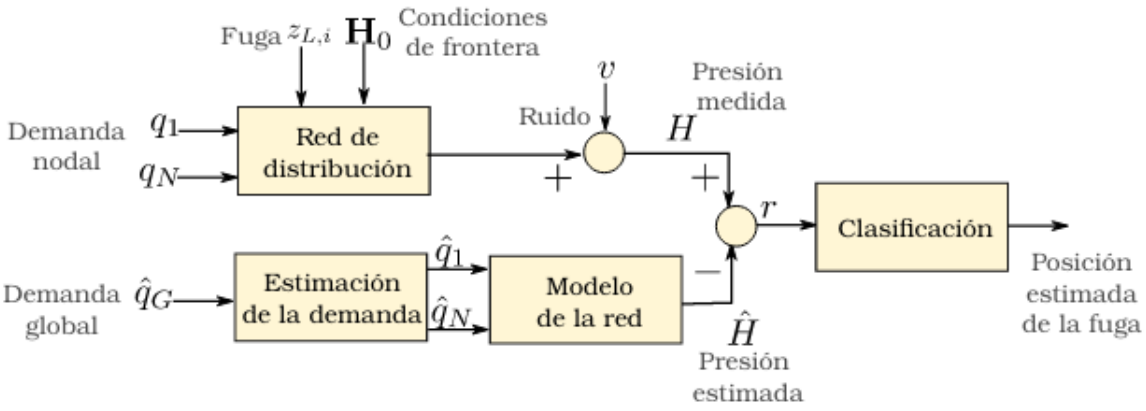

Figura 1. Esquema general de localización de fugas. Fuente: Elaboración propia.

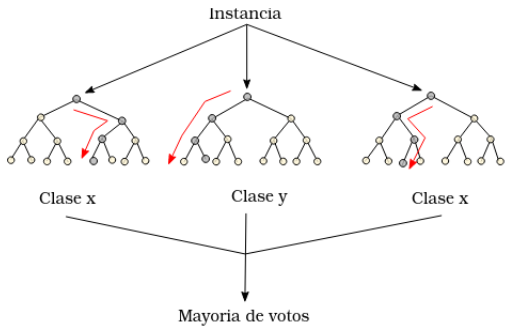

Figura 3. Bosque aleatorio. Fuente: Elaboración propia.

El diseño de un árbol de decisión requiere la selección de un medida de selección de atributos, existen diferentes enfoques para esta tarea donde en la mayoría se asignan una medida de calidad directamente al atributo. Una medida de selección de atributos frecuentemente utilizada es el criterio de la relación de ganancia de información. Cuando se utiliza el índice de ganancia de información, se asocia una medida de selección de atributo, que mide la impureza de un atributo con respecto a las clases. Para un conjunto de entrenamiento determinado Tr, seleccionando al azar y considerando que pertenece a alguna clase C_i, el índice de ganancia se puede escribir como:

$$\sum \sum_{j\neq} (f(C_i T_r)/|T_r|)(f(C_i, T_r)/|T_r|) \qquad (16)$$

donde $f(C_i, T_r)/|T_r|$ es la probabilidad de que el caso seleccionado pertenezca a la clase C_i. El número de características utilizadas en cada nodo para generar un árbol y el número de árboles son parámetros definidos por el usuario necesarios para generar un clasificador de bosque aleatorio [Pal, 2005].

II-E. Métricas de desempeño

El desempeño de los clasificadores se medir a través de diferentes métricas de desempeño. Para definir las métricas es necesario definir los siguientes conceptos: Un **verdadero positivo** es un resultado en el que el modelo predice correctamente la clase positiva. De manera similar, un **verdadero negativo** es un resultado en el que el modelo predice correctamente la clase negativa. Por otro lado, un **falso positivo** es un resultado en el que el modelo predice incorrectamente la clase positiva y finalmente un **falso negativo** es un resultado en el que el modelo predice incorrectamente la clase negativa. En este caso la clase positiva es la ocurrencia de una fuga en un determinado nodo y ninguna fuga es una clase negativa. Considerando las definiciones anteriores se definen las siguientes métricas de desempeño.

- La precisión: se define como el número de verdaderos positivos (T_p) sobre el número de verdaderos positivos más el número de falsos positivos (Fn):

$$\text{Precisión} = \frac{T_p}{T_p + F_p}$$

- Sensibilidad: se define como el número de verdaderos positivos sobre el número de verdaderos positivos màs el número de falsos negativos:

$$\text{Sensibilidad} = \frac{T_p}{T_p + F_n}$$

- Índice de desempeño F1: considera tanto la precisión como la sensibilidad de la prueba para calcular la puntuación:

$$\text{ìndice F1} = 2 \times \frac{\text{Precisión} \times \text{Sensibilidad}}{\text{Precisión} + \text{Sensibilidad}}$$

III. Localización de fugas

El desarrollo de las técnicas de clasificación requiere un previo entrenamiento fuera de línea que tiene por

objetivo obtener clasificador que funcione bajo escenarios con mediciones ruidosas. El procedimiento para emplear las técnicas de clasificación en un escenario real comienza con la calibración del modelo, tratando de obtener una descripción lo más realista posible [Ostfeld et al., 2008], para después generar una librería de faltas que contenga información para cada potencial falta bajo con diferentes magnitudes y condiciones incertidumbre [Soldevila et al., 2017]. La etapa de generación de datos es fundamental ya que la disponibilidad de datos representativos es una condición necesaria para obtener un buen clasificador. Dado que los datos que se pueden obtener de la red real pueden ser limitados, una forma de obtener un conjunto completo de datos de entrenamiento es mediante el uso del simulador hidráulico [Soldevila et al., 2017]. Bajo lo las limitantes para la obtención de datos en una red real, se emplea un esquema (mostrado en la figura Figura 4) para la generación de datos, donde la red real es sustituida por una simulación perturbada con ruido.

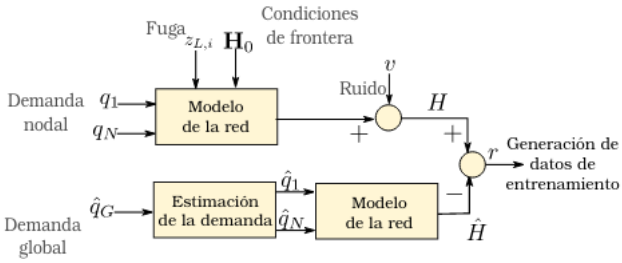

Figura 4. Esquema de generación de datos. Fuente: Elaboración propia.

En este caso para el análisis, simulación y evaluación del algoritmo de posicionamiento, se considera un problema tipo y la herramienta de simulación WNTR (del inglés: Water Network Tool for Resilience) desarrollada por la agencia de protección Ambiental de Estados Unidos (US EPA, por sus siglas en inglés). para entornos de simulación de lenguaje Python. El problema tipo utilizado corresponde a una de las redes utilizada como parte de la competición "Batlle of the Water Sensor Networks"(BWSN) [Ostfeld et al., 2008], la cual se compone de 178 tubos, 126 nodos, 2 tanques y un reservorio (ver Figura 5). Para la simulación de la red se utiliza el modelo de demanda en función de la presión, presentado anteriormente, en el cual se utiliza el algoritmo de Newton-Raphson para la aproximación simultanea del gasto y la presión. La posición y el numero de sensores se define siguiendo la metodología presentada en la sección II-B y en [Sarrate et al., 2014], el esquema resultante consta de tres sensores posicionados en los nodos 22, 66, 113 los cuales se presentan en la Figura 5 por medio de estrellas:

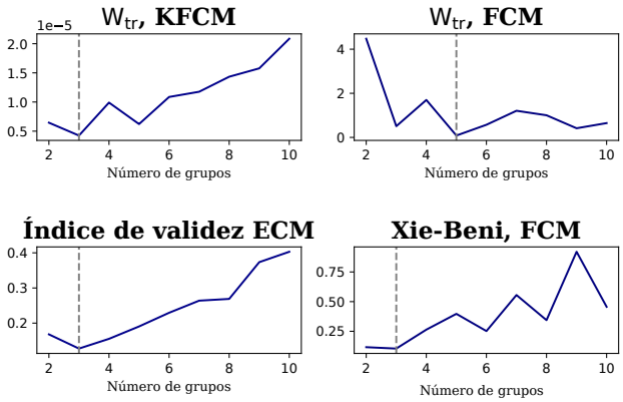

Figura 6. Indices de validez en grupos. Fuente: Elaboración propia.

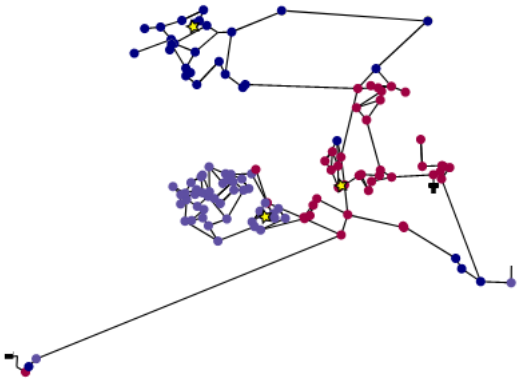

Figura 5. Posición y número de sensores, red BWSN. Fuente: Elaboración propia.

Para definir un número reducido de sensores a instalar en los posibles 126 nodos, se evalúan las técnicas de agrupamiento descritas en la es decir FCM, KFCM y ECM, donde el número adecuado de grupos se define con información de una matriz de sensibilidad y los correspondientes índices de validación de las técnicas de agrupamiento. Los resultados de estos índices para una evaluación de 1 a 10 potenciales grupos se muestran en la Figura 6,los cuales coinciden en tres sensores, únicamente en el índice W_tr evaluado en el algoritmo FCM, el resultado es cinco grupos. Debido al número de sensores obtenido en la mayoría de los índices, se considera tres como el número de grupos y sensores a instalar.

Unas vez definidos los grupos, la posición de los sensores es determinada por medio de un algoritmo genético usado para abordar el problema de minimización basado en proyecciones, descrito en la sección II-B En ese sentido, para la ejecución del algoritmo genético se consideró una población de 30 cromosomas, una ejecución iterativa de 100 ejecuciones y un crossover

de recombinación de 0.7.

Para el entrenamiento se realizaron simulaciones generando datos para todas las posibles posiciones fuga, bajo diferentes periodos de simulación, para ello se utilizaron 1000 ejemplos diferentes de fugas. Aunado a lo anterior, también se generaron datos con diferentes tamaños de orificio en un rango comprendido entre 0.01m y 0.07m en cada tubería de la red, considerando que el diámetro de las tuberías bajo estudio se encuentran entre un rango de 0.17 y 0.60 m. para el clasificador de bosque aleatorio se consideraron 100 árboles.

IV. Resultados

Para evaluar el desempeño del algoritmo se comparo con dos clasificadores más basado en el algoritmo de vecinos cercanos (KNN) y máquinas de vector de Soporte (SVM). Para el clasificador basado en KNN se consideró un rango de cuatro vecinos cercanos y una métrica euclidiana, mientras que para el clasificador basado SVM se utilizo un kernel con una función gaussiana con parámetro de penalización para el error de 1×10^7. Los resultado del en entrenamiento alcanzados por cada clasificador se resumen mediante la Tabla I de acuerdo con tres métricas de desempeño.

Tabla I
Métricas de desempeño e los clasificadores en el entrenamiento. Fuente: Elaboración propia.

Clasificador	Precisión	Sensibilidad	F1
KNN	0.55	0.56	0.51
SVM	0.71	0.66	0.66
RF	0.88	0.88	0.88

Para ilustrar los resultados del entrenamiento en la Figura 7 se muestra el resultado de localización obtenido por cada uno los clasificadores en tres ejemplos, donde el punto rojo denota la posición real de la fuga.

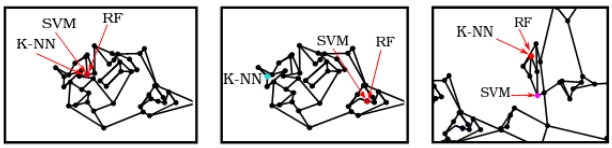

Figura 7. Predicciones de ejemplo durante el entrenamiento. Fuente: Elaboración propia.

Los resultados del entrenamiento permiten observar que el mayor desempeño lo obtuvo el clasificador RF seguido del SVM y finalmente el algoritmo KNN. Una vez entrenado cada algoritmo se realizaron nuevas pruebas en 250 escenarios de fugas. La Tabla II muestra resultados para 10 de los 250 ejemplos, mientras que la Figura 8 muestra un ejemplo de las señales de los sensores 22, 66 y 133, en los que se considera que tiene un sensor instalado, para la fuga simulada en nodo 84, donde la ocurrencia de la fuga ocurre en la hora treinta marcada con la linea roja).

Tabla II
Muestra de localización de fugas por los diferentes clasificadores. Fuente: Elaboración propia.

Posición simulada	Predicciones		
	KNN	SVM	RF
84	84	84	84
40	45	40	40
52	52	50	52
108	108	4	4
71	72	71	72
122	125	125	125
30	22	41	22
109	95	94	94
58	59	59	64
12	39	39	10
22	27	26	27

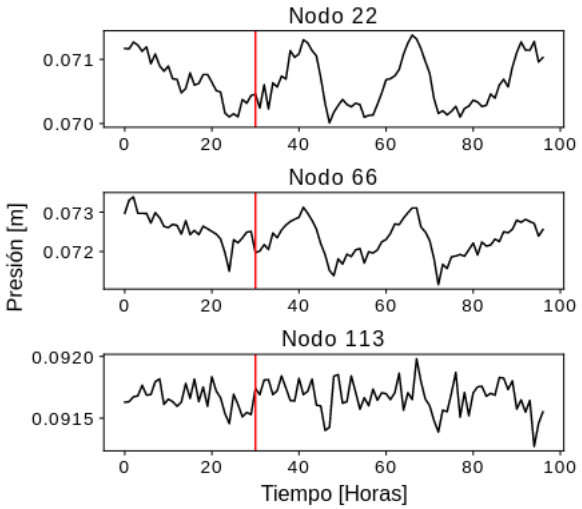

Figura 8. Ejemplo de señales de los sensores 22, 66 y 113 en presencia de una fuga. Fuente: Elaboración propia.

En la Tabla III se pude observar una disminución en las métricas de desempeño para los 250 casos, sin embargo, de la Tabla II se puede observar que las predicciones se mantiene en la vecindad del valor real.

Tabla III
Métricas de desempeño e los clasificadores. Fuente: Elaboración propia.

Clasificador	Precisión	Sensibilidad	F1
KNN	0.11	0.12	0.10
SVM	0.14	0.14	0.13
RF	0.15	0.14	0.13

V. Conclusiones

En este artículo se evaluó un clasificador basado en bosque aleatorio para abordar el problema de localización de fugas en redes de distribución y se comparo con dos

diferentes algoritmos. Los resultados muestran que el esquema propuesto mejora el desempeño alcanzado por el clasificador con respecto a los otros clasificadores en un caso de estudio, aun cuando la información de los sensores está perturbada por ruido. Una de las tendencias actuales encontradas en la literatura es la fusión de clasificadores, donde se toma una decisión consensuada de varios clasificadores la cual tiende a reducir el error. Esta filosofía puede resultar útil para abordar el problema de localización de fugas.

REFERENCIAS

[A Miyamoto et al., 2008] A Miyamoto, S., Ichihashi, H., and Honda, K. (2008). *Algorithms for Fuzzy Clustering Methods in c-Means Clustering with Applications*. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Cabrera and Vela, 2013] Cabrera, E. and Vela, A. F. (2013). *Improving Efficiency and Reliability in Water Distribution Systems*. Springer Science Bussines Media, B.V., Heidelberg.

[Casillas et al., 2015] Casillas, M. V., Garza-Castañón, L. E., and Puig, V. (2015). Optimal Sensor Placement for Leak Location in Water Distribution Networks using evolutionary algorithms. *Water*, 7:6496–6515.

[Gamboa-Medina and Reis, 2017] Gamboa-Medina, M. M. and Reis, L. F. R. (2017). Sampling Design for Leak Detection in Water Distribution Networks. In *Procedia Engineering*, pages 460–469, Cartagena.

[Girolami, 2002] Girolami, M. (2002). Mercer kernel-based clustering in feature space. *IEEE Transactions on Neural Networks*.

[Giustolisi et al., 2008] Giustolisi, O., Savic, D., and Kapelan, Z. (2008). Pressure-Driven Demand and Leakage Simulation for Water Distribution Networks. *Journal of Hydraulic Engineering*, 134(5):626–635.

[Klise et al., 2017] Klise, K., Hart, D., Moriarty, D., Washington, M. B., DC, U., and 2017, U. (2017). Water network tool for resilience (WNTR) user manual. Technical report, United States Environmental Protection Agency.

[Leu and Bui, 2016] Leu, S. S. and Bui, Q. N. (2016). Leak Prediction Model for Water Distribution Networks Created Using a Bayesian Network Learning Approach. *Water Resources Management*.

[Masson and Denœux, 2008] Masson, M. H. and Denœux, T. (2008). ECM: An Evidential Version of the Fuzzy C-Means Algorithm. *Pattern Recognition*, 41(4):1384–1397.

[Ostfeld et al., 2008] Ostfeld, A., Uber, J. G., Salomons, E., Berry, J. W., Hart, W. E., Phillips, C. A., Watson, J.-P., Dorini, G., Jonkergouw, P., Kapelan, Z., et al. (2008). The battle of the water sensor networks (bwsn): A design challenge for engineers and algorithms. *Journal of Water Resources Planning and Management*, 134(6):556–568.

[Pal, 2005] Pal, M. (2005). Random forest classifier for remote sensing classification. *International journal of remote sensing*, 26(1):217–222.

[Pudar and Liggett, 1992] Pudar, R. S. and Liggett, J. A. (1992). Leaks in Pipe Networks. *Journal of Hydraulic Engineering*, 118(7):1031–1046.

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

[Rossman, 2001] Rossman, L. A. (2001). Epanet 2 users manual. *US Environmental Protection Agency, Cincinnati, Ohio*.

[Sarrate et al., 2014] Sarrate, R., Blesa, J., and Nejjari, F. (2014). Clustering Techniques Applied to Sensor Placement for Leak Detection and Location in Water Distribution Networks. In *22nd Mediterranean Conference on Control and Automation, 2014*, pages 109–114, Palermo. IEEE.

[Soldevila et al., 2017] Soldevila, A., Fernandez-Canti, R. M., Blesa, J., Tornil-Sin, S., and Puig, V. (2017). Leak localization in water distribution networks using Bayesian classifiers. *Journal of Process Control*.

[Todini, 2010] Todini, E. (2010). A more realistic approach to the "extended period simulation" of water distribution networks. In *Advances in Water Supply Management*. Taylor & Francis.

[Valizadeh et al., 2009] Valizadeh, S., Moshiri, B., and Salahshoor, K. (2009). Leak detection in transportation pipelines using feature extraction and KNN classification. In *Pipelines 2009: Infrastructure's Hidden Assets - Proceedings of the Pipelines 2009 Conference*.

[World Water Assessment Programme United Nations and UN-Water, 2009] World Water Assessment Programme United Nations and UN-Water (2009). *Water in a changing world*, volume 1. Earthscan.