Estado del arte de algoritmos de Machine Learning para la detección de rupturas súbitas

Jaime Ernesto Chiang Cruz E-MAIL: jaime.cc@civil.cujae.edu.cu
Centro de Investigaciones Hidráulicas, Universidad Tecnológica de La Habana José Antonio Echeverría, CUJAE,Cuba.

Iliover Vega González E-MAIL: ivegag@cime.cujae.edu.cu
CIME, Universidad Tecnológica de La Habana José Antonio Echeverría, CUJAE, Cuba.

Jorge Ramírez Beltrán E-MAIL: jorgeramirezcihcujae@gmail.com
Centro de Investigaciones Hidráulicas, Universidad Tecnológica de La Habana José Antonio Echeverría, CUJAE,Cuba.

ABSTRACT

En este trabajo se realiza una revisión de los paradigmas existentes y las técnicas más usadas en la detección de rupturas súbitas, profundizando en las que emplean Machine Learning como herramienta principal para la interpretación de datos. Se comparan la relación entre la efectividad de la detección y los parámetros de cada algoritmo, así como el nivel de procesamiento requerido. Para la Máquina de Soporte Vectorial la efectividad en la detección de ruptura súbita está relacionada de forma exponencial con la cantidad de combinaciones de C y γ . El árbol de decisión expuesto aumenta su precisión mientras más información del estado de la red tenga. La red neuronal analizada demuestra una efectividad en la detección al nivel del resto de algoritmos tratados manteniendo el compromiso con el nivel de procesamiento.

PALABRAS CLAVES: Aprendizaje automático, redes hidráulicas, redes neuronales, ruptura súbita.

State of the art of Machine Learning algorithms for burst detection

RESUMEN

In this work, a review of the existing paradigms and the most used techniques in the burst detection is carried out, delving into those that use Machine Learning as the main tool for data interpretation. The relationship between detection effectiveness and the parameters of each algorithm, as well as the level of processing required, are compared. For Support Vector Machine, the effectiveness in burst detection is exponentially related to the number of combinations of C and γ. The exposed decision tree increases its precision the more information about the state of the network it has. The artificial neural network demonstrates a detection effectiveness at the level of the rest of the algorithms treated, maintaining the commitment to the processing level.

KEYWORDS: Machine learning, hydraulics networks, neural networks, burst.

CITMA CERTIFICADO

Junio 2024-RECIBIDO *APROBADO-Junio 2024*

01 INTRODUCCION

El agua es un recurso de vital importancia, contribuye a la estabilidad y regulación de los entornos, del medio ambiente y de los organismos que habitan en este. El riesgo de agotar los recursos hidráulicos se ha incrementado notablemente en los últimos años debido a la creciente contaminación y el cambio climático los cuales se consideran un peligro inminente. La mayor influencia para las pérdidas de agua en los sistemas hidráulicos está dada por rupturas súbitas y fugas de fondo. Las primeras son grandes fugas de agua que ocurren en un corto intervalo de tiempo, ocasionadas por rupturas bruscas en los conductos y componentes de su ensamblaje. Las fugas de fondo presentan un caudal bajo y principalmente se originan en tanques y tuberías. A pesar de que estas últimas contribuyen a la pérdida de agua, la ruptura súbita representa un volumen mayor.

Por la importancia que presenta la detección oportuna de rupturas súbitas en redes hidráulicas, es esencial la búsqueda de alternativas que mejoren su eficiencia y lleven a cabo una gestión adecuada de los recursos hidráulicos. Una de ellas es tener un sistema capaz de detectar fugas por ruptura súbita en tiempo real, que reduzca al mínimo los falsos positivos y logre la suficiente precisión al conseguir un manejo rápido y efectivo del problema. Este sistema requiere de un algoritmo que cumpla con las exigencias anteriores y pueda ser implementado en una plataforma de bajo consumo de potencia. Esto posibilita que los registradores de datos se mantengan autónomos durante un período de tiempo prolongado, facilitando su ubicación en lugares de difícil acceso.

En la comunidad científica se definen dos paradigmas compuestos por técnicas que detectan y localizan estos fenómenos indeseados. El primero adquiere datos a través de los sistemas de Supervisión, Control y Adquisición de Datos (SCADA, por sus siglas en inglés de Supervision Control and Data Adquisition) con un período de muestreo que típicamente se encuentra entre 5 y 15 minutos, debido a que este período de muestreo provee un balance razonable entre el volumen de datos y la definición de patrones diarios, también permite obtener una buena representación de la dinámica del fluido en la red (Trutié-Carrero et al. 2018). A pesar de los resultados alcanzados siguiendo el modelo mencionado, la principal dificultad es su incapacidad para detectar y localizar la ruptura súbita en el momento que se genera; no disminuyendo su ciclo de vida y aumentando los costos. El segundo paradigma es el basado en datos (data-driven), este tiene un período de muestreo mucho más pequeño para lograr una mejor apreciación del instante en el que se generó la ruptura súbita, utilizando en la adquisición de los datos, tecnología de sensores inalámbricos desplegada en el Sistema de Distribución de Agua (SDA) con dispositivos de bajo costo, posibilitando monitorizar la infraestructura en tiempo real (Srirangarajan et al., 2013). En los métodos basados en datos no es necesario un conocimiento profundo del SDA, pues solo implican un análisis estadístico o de procesamiento de señales de los datos adquiridos, como consecuencia sus resultados son resistentes a errores de modelado y medición.

En los últimos años el desarrollo del Machine Learning se ha visto potenciado. Es una rama de la Inteligencia Artificial que desarrolla técnicas que permiten a máquinas tomar decisiones, aprenden gracias a encontrar patrones en los datos y predecir situaciones posteriores. En este contexto, el presente trabajo tiene como objetivo exponer los paradigmas existentes y las técnicas más usadas en la detección de rupturas súbitas, profundizando en las que emplean Machine Learning como herramienta para interpretar los datos de presión y caudal.

02 DESARROLLO

DETECCIÓN DE RUPTURAS SÚBITAS EN TUBERÍAS

Las fugas por ruptura súbita en los SDA representan un volumen mayor de pérdida de agua que las de fondo, provocando un aumento en la perturbación del sistema. Además de que produce la interrupción del suministro de agua, puede dañar las calles y edificaciones y eleva los costos por reparación.

En la figura 1 se observan los tiempos que intervienen en el ciclo de vida de dicho evento.

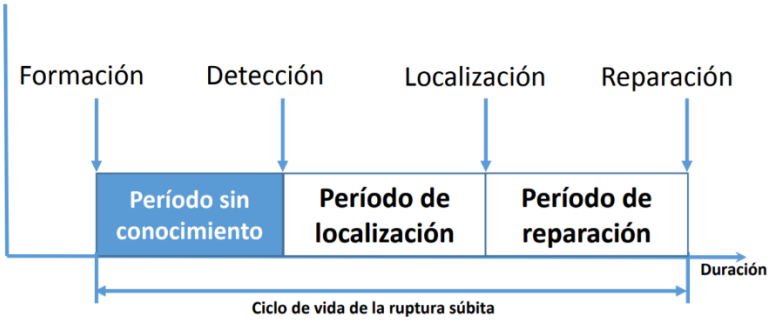

Figura 1. Ciclo de vida de la ruptura súbita.

En la mayoría de los casos, una técnica para la detección de rupturas súbitas basada en datos se realiza siguiendo dos pasos:

Adquisición, preprocesamiento y transformación de datos: Su objetivo principal es eliminar datos erróneos o faltantes de los adquiridos en la medición de parámetros de interés en la tubería como caudal o presión, para facilitar los análisis posteriores. Estos datos deben procesarse antes de ser usados para el análisis de fugas, lo que asegura que son adecuados para el algoritmo seleccionado. Aunque es posible que cuestiones relacionadas con la variabilidad y la incertidumbre surjan durante el procesamiento de los datos de medición, pueden ser superadas en cierta medida con preprocesamiento y transformación de los datos (Zaman et al. 2020).

Estrategia de detección de fugas: Las técnicas basadas en datos se pueden clasificar según sus estrategias de detección de fugas en métodos de clasificación de conjuntos de características, métodos de clasificación-predicción y métodos estadísticos.

ALGORITMOS DE MACHINE LEARNING

Machine Learning es una disciplina de la Inteligencia Artificial cuyo objetivo principal es que un agente artificial mejore su rendimiento a partir de resultados previamente obtenidos. Desde el punto de vista del Machine Learning obtener conocimiento implica extraer información potencialmente útil y desconocida a partir de un conjunto de datos. Para lograrlo, es necesario generar métodos, generalmente no-triviales, que permitan analizar los datos, y a partir de dicho análisis producir un modelo mediante el cual pueda representar un fenómeno. En Machine Learning se pueden diferenciar dos tipos de métodos que se utilizan para la construcción de los modelos previamente mencionados: el aprendizaje supervisado y no supervisado.

El aprendizaje no supervisado se basa en modelos descriptivos en donde la máquina entiende los datos, la evaluación es cualitativa o indirecta y no realiza predicciones, encuentra algo específico. El aprendizaje no supervisado utiliza algoritmos de aprendizaje automático para analizar y agrupar un conjunto de datos sin etiquetar. Estos algoritmos hallan patrones ocultos en los datos sin la necesidad de ser supervisados.

En el aprendizaje supervisado se utilizan modelos predictivos en donde la máquina aprende explícitamente, por lo tanto, intenta predecir el futuro a partir de datos históricos, además resuelve problemas de clasificación y regresión. A medida que los datos de entrada se introducen en el modelo, este ajusta sus ponderaciones hasta que el modelo se ha ajustado correctamente, lo que ocurre como parte del proceso de validación cruzada (Saravanan y Sujatha, 2018).

Existen diversas técnicas que permiten obtener modelos que representen el patrón obtenido a partir de los conjuntos de datos. Los árboles de decisión, las redes neuronales y las máquinas de soporte vectorial han sido ampliamente utilizados para construir modelos predictivos. Resulta interesante entonces profundizar en las características de estos algoritmos.

ÁRBOL DE DECISIÓN

Los Árboles de Decisión (DT por sus siglas del inglés Decision Trees) son algoritmos estadísticos o técnicas de Machine Learning que permiten la construcción de modelos predictivos de analítica de datos basados en su clasificación según ciertas características o propiedades, o en la regresión mediante la relación entre distintas variables para predecir el valor de otra. Por otro lado, los árboles de decisiones brindan una eficiencia de alto nivel y una fácil interpretación. Estos dos beneficios hacen que este simple algoritmo sea popular en el espacio del aprendizaje automático.

Los árboles de decisión se clasifican en dos tipos de modelos: regresión y clasificación (Mendoza García, 2023).

Modelos de Regresión: se pretende predecir el valor de una variable en función de otras variables que son independientes entre sí. Por ejemplo, predecir el precio de venta de un piso en función de variables como su localización, superficie, distancia a la playa, etc. El posible resultado no forma parte de un conjunto predefinido, sino que puede tomar cualquier posible valor.

Modelos de Clasificación: se pretende predecir el valor de una variable mediante la clasificación de los datos en función de otras variables. Por ejemplo, predecir qué personas comprarán un determinado producto, clasificando entre clientes y no clientes, o qué marcas de computadoras portátiles comprará cada persona mediante la clasificación entre las distintas marcas. Los valores para predecir son predefinidos, es decir, los resultados están definidos en un conjunto de posibles valores.

En el artículo (Alves Coelho et al. 2020) se presenta un método de detección de rupturas súbitas que emplea árboles de decisión para interpretar cambios rápidos en las variaciones transitorias de una señal de presión. En el algoritmo presentado se utiliza una partición jerárquica de datos de entrenamiento y un atributo determinado para dividir los datos, y esta división se realiza de forma iterativa hasta que el nodo hoja contenga una serie de registros que se pueden utilizar para clasificar los datos. Cada nodo en el árbol actúa como un caso de prueba para algún atributo, y cada borde que desciende de ese nodo corresponde a una de las posibles respuestas al caso de prueba. Este proceso es recursivo y se repite para cada subárbol enraizado en los nuevos nodos.

Para la elección del atributo se utilizan Medidas de selección de atributos (MSA) para dividir los registros. Las MSA consisten en seleccionar el subconjunto más pequeño de atributos tal que no se afecte significativamente el porcentaje de clasificación y que la distribución de clases resultante sea lo más parecido posible a la original.

El proceso de selección de atributos involucra 4 pasos (Arredondo Arteaga et al. 2017):

1. Generación de candidatos (subconjuntos): involucra una estrategia de búsqueda
2. Evaluación de candidatos (subconjuntos): requiere un criterio de evaluación

3. Criterio de parada: Este puede darse por la estrategia de búsqueda, el número de iteraciones realizadas, el número de atributos seleccionados, el que no se mejore el criterio de evaluación al añadir/quitar otro atributo, que el error de clasificación esté por debajo de un umbral, etc.
4. Validación de resultados: Ya sea contra atributos conocidos como relevantes o comparando el error en la clasificación con y sin la selección de atributos.

Se utiliza el algoritmo ID3, el cual induce árboles de decisión basados en la ganancia de información obtenida de los ejemplos de entrenamiento y luego lo usa para clasificar el conjunto de prueba. En la figura 2 se muestra el pseudocódigo del algoritmo.

1: **procedimiento** ID3(R, C, S) ▷ Entrada: R un conjunto de atributos no objetivo; C, el atributo objetivo; S, dato de entrenamiento. Salida: árbol de decisión
2: **si** $S = \emptyset$ **entonces**
3: **regresar** $\emptyset$ ▷ Regresar conjunto vacío
4: **fin si**
5: **si** S tiene valores con el mismo valor objetivo **entonces**
6: **regresar** un único nodo con este valor
7: **fin si**
8: $D \leftarrow Calcular\ la\ Ganancia(D, S) con\ 2.5$
9: $\{d_j j = 1, 2, 3..., m\} \leftarrow valores\ de\ atributo\ de\ D$
10: $\{S_j \quad con \quad j \quad = \quad 1, 2, 3..., m\} \quad \leftarrow$ *los subconjuntos de E construído de* d_j *registrar el valor del atributo* D
11: **regresar** Un árbol cuya raíz es D y los arcos son etiquetados por $d_1, d_2, \cdots, d_m$. E ir a los sub-árboles ID3$(R - \{D\}, C, S_1), ID3(R - \{D\}, C, S_2), \cdots, ID3(R - \{D\}, C, S_m)$
12: **fin procedimiento**

Figura 2. Pseudocódigo del algoritmo ID3 (Rodríguez et al. 2021)

REDES NEURONALES ARTIFICIALES

Las Redes Neuronales Artificiales (RNA) están inspiradas en la biología, esto significa que están formadas por elementos que se comportan de manera análoga a las neuronas definidos con el mismo nombre (en las funciones más elementales) y están organizadas de una forma similar a la del cerebro. Cuando se trabaja con grandes cantidades de neuronas, es natural ordenar aquellas que tienen comportamientos similares en "capas". Cada capa es un vector de neuronas. Se puede establecer una clasificación de las redes neuronales en cuanto al número y características de sus capas como simples o multicapa, como se muestra en la figura 3.

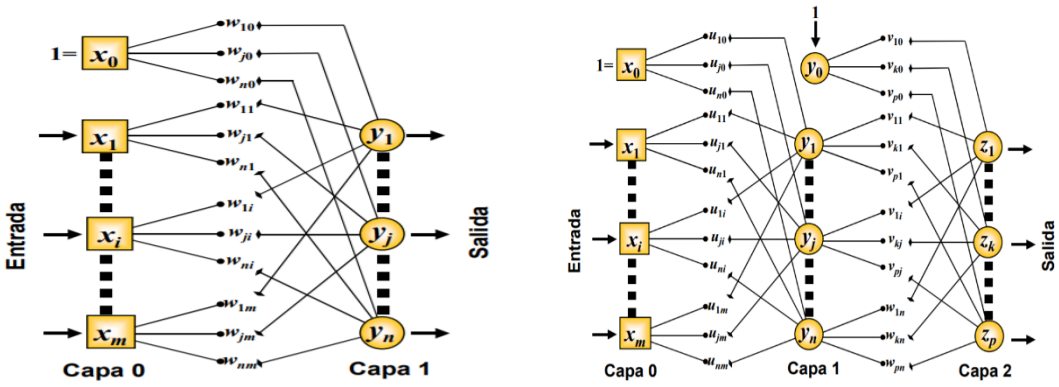

Figura 3. A la izquierda la representación gráfica de una red de capa simple y a la derecha representación gráfica de una multicapa (Hernández, 2014)

En (Bohorquez et al., 2021) se presenta una técnica basada en el análisis de variaciones transitorias de presión que emplea RNA para interpretar cambios rápidos de la señal para identificar, localizar y caracterizar rupturas súbitas en tuberías de agua. El enfoque propuesto se basa en datos y no en información detallada sobre la tubería analizada para la interpretación de la traza de presión transitoria casi en tiempo real. Se utiliza la RNA como herramienta para interpretar los rastros e identificar ráfagas. Esta técnica analiza segmentos de tiempo corto de datos de la señal potencialmente continuo a través de dos procesos diferentes. Un primer análisis para detectar si una ventana de tiempo de la variación transitoria de presión en particular contiene cambios anormales que podrían haber sido producidos por la ocurrencia de una ruptura súbita. Este análisis inicial también determina si la información contenida en la ventana de tiempo es suficiente para localizar y caracterizar el estallido potencial. Luego, una segunda etapa utiliza otra RNA para analizar las ventanas de tiempo anormales detectadas desde la primera y lograr predecir la ubicación y las características de la ráfaga que se produce.

La estructura de RNA más general y ampliamente utilizada no es capaz de capturar adecuadamente los cambios en la presión debido a la presencia de diferentes elementos en una tubería (Bohorquez et al. 2021). Teniendo esto en cuenta, es mejor seleccionar una arquitectura de red convolucional unidimensional (1D), ya que estas redes son más ligeras y menos propensas al sobreajuste. Están diseñadas para funcionar satisfactoriamente tanto en la detección de la ocurrencia de una ráfaga como en la identificación de la ubicación y el tamaño de la ráfaga. El diseño se desarrolla modificando diferentes características de la arquitectura, incluido el número de capas convolucionales, el tipo de función de activación, el número de filtros en cada capa y el tamaño del lote de entrenamiento.

La configuración final de las redes convolucionales 1D incluye un máximo de siete capas convolucionales, el uso de unidad lineal rectificada con fugas (Leaky ReLU) y Softmax como funciones de activación, un número máximo de 12 filtros que incrementa en cada capa convolucional, un tamaño de lote de entrenamiento de 50 muestras, y tres capas densas de tamaños máximos de 21, 9 y 3.

Se requiere un procesamiento adicional de la variación transitoria de presión porque la técnica propuesta está diseñada para funcionar casi en tiempo real. Se aplica el concepto de ventana de tiempo deslizante (Mounce et al. 2011). Una vez que se completa la partición de la señal, las ventanas de tiempo resultantes se pueden clasificar en tres categorías dependiendo de la información de cabeza contenida.

La primera categoría se define como condición de traza normal, o Categoría N, ya que estas ventanas de tiempo solo contienen parte de la supuesta variación de la sinusoidal lenta antes de la ruptura súbita. La segunda categoría se define como condición anormal de la traza con información incompleta para la identificación o Categoría Ab-I. Las ventanas de tiempo en esta categoría capturan la caída de carga inicial debido al estallido, pero no se incluye el reflejo de la onda de estallido. La última categoría se define como condición anormal de la traza con información completa para la identificación, o Categoría Ab-C. Las ventanas de tiempo en esta categoría contienen un reflejo completo de la onda transitoria creada por el estallido en las condiciones de contorno de la tubería.

La duración de cada ventana de tiempo debe ser de al menos 2 segundos para capturar el primer conjunto de reflejos de onda causados por la ocurrencia de la ruptura súbita (P. Huang et al., 2018). Se requiere un transductor de presión para capturar las variaciones de este parámetro en tiempo real, se toman 250 muestras por segundo que están en formatos enteros de 16 bits. Teniendo en cuenta la frecuencia de muestreo y la duración de cada ventana de tiempo, cada traza se transformó en 1328 ventanas de tiempo cada una con una longitud de 562 valores (correspondientes a 2,25 s).

Cada uno de estos valores debe ser almacenado en RAM durante la ejecución del algoritmo para su procesamiento, para lo que se necesitaría ocupar 1,124 Kb de memoria.

MÁQUINA DE SOPORTE VECTORIAL

En (Liu et al. 2019) se propone un método de identificación de fugas basado en una SVM. Con base en las diferencias en las características de tiempo-frecuencia de las señales con fugas y sin fugas, se propone un método que construye matrices de características empleando la función de densidad de espectro, la entropía aproximada y el análisis de componentes principales (PCA) y que utiliza SVM como un clasificador para identificar las fugas. Las pruebas para la comprobación de la efectividad en la detección de rupturas súbitas de este algoritmo se realizan a lo largo de un tubo compuesto de aluminio y plástico expuesto, con un diámetro de 27 mm.

Característica de densidad de espectro

Se ha comprobado que los componentes de los espectros de las señales de fuga se concentran principalmente en bandas específicas. Como resultado, las diferencias entre los espectros de las señales pueden emplearse como características de identificación de fugas en tuberías. Para extraer las diferencias de la densidad espectral de la señal y realizar un análisis de tiempo-frecuencia de las señales de la tubería, se utiliza la descomposición de modo empírico (EMD por sus siglas en inglés de Empiric Mode Decomposition) y una función de dominio de frecuencia para la detección. EMD puede descomponer selectivamente la señal como la suma de un número finito de funciones de modo intrínseco (IMF por sus siglas en inglés de Intrinsic Mode Function).

La figura 4 muestra el espectro de los primeros cuatro grupos de componentes IMF que se obtienen a partir de la EMD de la señal de fuga y de no fuga de la tubería. Comparando el IMF de cada capa de la señal de fuga y sin fuga, se puede ver que el espectro principal de la señal estaba en el IMF de la primera capa. El espectro de la señal de fuga en esta capa se distribuye principalmente entre los 1000 Hz y 2000 Hz. El espectro de la señal sin fugas es más aleatorio y se distribuyó principalmente en toda la banda.

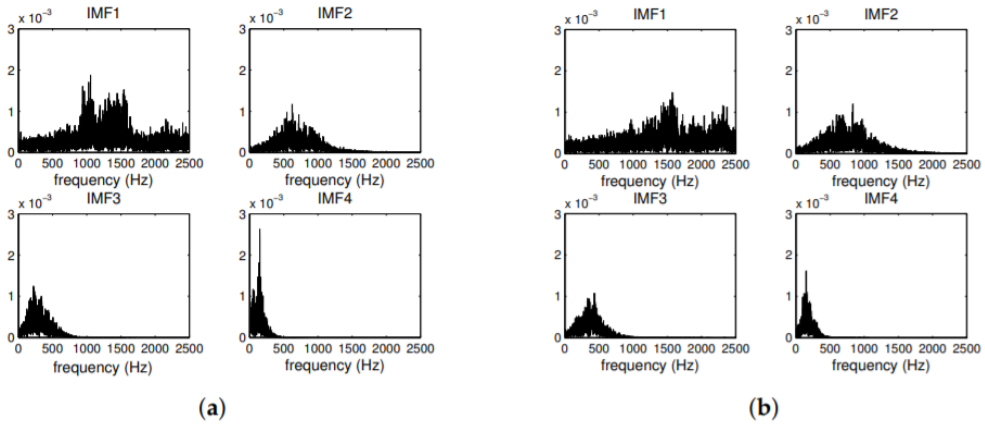

Figura 4. Los primeros cuatro espectros de función de modo intrínseco (IMF) de la señal de fuga y no fuga de la tubería. (a) Señal de fuga. (b) Señal sin fugas (Van Jaarsveldt et al. 2023)

Característica de complejidad de la señal

Debido a que la fuga en la tubería es un evento localizado y de baja probabilidad de ocurrencia, debe haber diferencias en la composición en el dominio del tiempo de las señales de fuga y las señales sin fuga, la composición de las señales de fuga debería ser más compleja. La entropía aproximada (ApEn por sus siglas del inglés de Aproximate Entropy) es la probabilidad condicional cuando se

mantiene la similitud después de que las dimensiones de un vector de similitud aumentan de m a m + 1 y es la probabilidad del nuevo modo cuando cambia el número de dimensiones.

Característica del componente principal de la señal

El análisis de las componentes principales (PCA por sus siglas en inglés de Principal Components Analise) es un método clásico de extracción de características que implica la reducción de la dimensionalidad y convierte las variables en un número menor de variables agregadas. Cada componente principal es una combinación lineal de las variables originales y los componentes principales individuales no están mutuamente correlacionados. Estos pueden transmitir la gran mayoría de la información contenida en las variables originales sin que se superponga entre sí.

Para aumentar la precisión de la detección de fugas, se aprovechan las características de tiempo-frecuencia para construir conjuntos de características de identificación usando SVM para clasificar las características de la señal y, por lo tanto, determinar la fuga de la tubería. El SVM es un medio ventajoso para resolver problemas de muestras pequeñas, problemas no lineales y problemas que involucran datos de alta dimensión.

Para mejorar la precisión de la detección de fugas, se deben usar muestras de entrenamiento y de prueba para optimizar el SVM. Debido a los efectos de los factores ambientales en las tuberías de agua soterrada, es necesario realizar señales de muestreo durante diferentes momentos y en diferentes lugares para compilar conjuntos de muestras que incluyen señales de fuga y señales de no fuga. Al principio, el conjunto de funciones de la muestra de entrenamiento se usa para realizar el entrenamiento de SVM, lo que crea un modelo de identificación preliminar. El conjunto de características de la muestra de prueba se usa luego para probar el modelo SVM entrenado. El modelo SVM se optimiza aún más en función de los resultados de las pruebas hasta que la precisión de la salida de la prueba cumpla con los requisitos, lo que da como resultado un modelo de identificación de fugas de tuberías SVM.

El análisis teórico del modelo de SVM indica que los principales factores que afectan su desempeño incluyen la función del núcleo γ y el factor de penalización. De acuerdo con las características y la capacidad del SVM, se toma el núcleo de base radial como la función kernel.

RENDIMIENTO DE LOS MÉTODOS DE DETECCIÓN DE RUPTURAS SÚBITAS

Para comprobar la efectividad de los algoritmos planteados se realizaron una serie de pruebas donde se utilizaron conjuntos de datos de señales de fuga y de no fuga, evaluándose la capacidad de cada uno en la detección de la ruptura súbita.

Con las RNA se logra clasificar correctamente la mayoría de las ventanas de tiempo, excepto una que pertenece a la Categoría N y fue clasificada como Categoría Ab-I, como se puede observar en la figura 5. Además, algunas otras ventanas de tiempo se clasificaron en la Categoría N, pero pertenecen a la Categoría Ab-I (lo que muestra un retraso en la capacidad de la RNA de detección de patrones para identificar la ocurrencia de una ruptura súbita), y dos ventanas de tiempo que pertenecen a la Categoría Ab-I se clasificaron en la Categoría Ab-C (no visible en el gráfico).

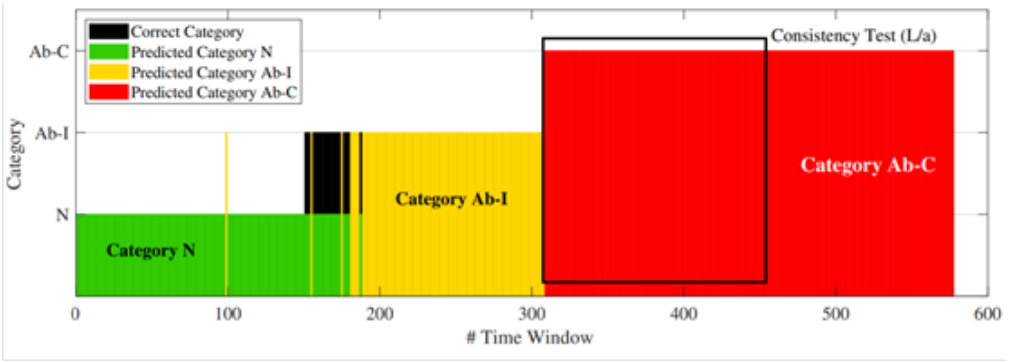

Figura 5. Clasificación de las ventanas de tiempo según la RNA (Bohorquez et al. 2021)

Las RNA de localización de rupturas súbitas brindan diferentes distribuciones de resultados, pero para los tres casos, la ubicación promedio pronosticada está dentro del 1,5 % de error. La predicción más cercana a la ubicación real de la ráfaga se obtuvo para la identificación de la tercera ráfaga RNA en la que se predijo la ubicación a solo 0,06 m de la ubicación real. De manera similar, la distribución en las predicciones del tamaño de la ruptura. Las tres RNA predijeron diferentes tamaños, lo que resultó en un error medio entre el 6 % y el 16 % del tamaño real. Aunque estos resultados son menos precisos que los resultados de ubicación, es importante señalar que el error medio más grande corresponde a un error absoluto de solo 0,32 mm. Además, el tiempo de cálculo necesario para obtener estos resultados fue de unos pocos segundos.

En el caso del modelo SVM su rendimiento está determinado por los parámetros C y γ. En (Liu et al. 2019) se extrajeron 50 conjuntos de datos de cada una de las señales de fuga y de no fuga y se usaron para crear un conjunto de entrenamiento. Las muestras restantes se usaron luego para crear un conjunto de prueba. Los parámetros SVM (C, γ) se establecieron en una potencia entera de dos; el rango de C se estableció como $C \in [2^{-5}, 215]$; y el rango de γ se estableció como $\gamma \in [2^{-15}, 25]$. Se usó el método de búsqueda en cuadrícula y se usaron 21 × 21 = 441 para combinaciones de (C, γ). La precisión de detección se muestra en la figura 6.

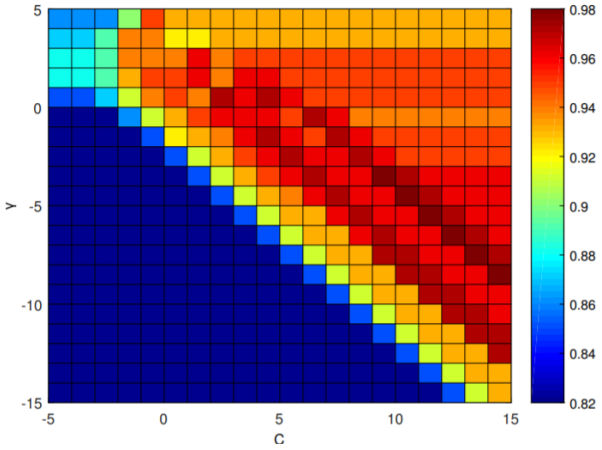

Figura 6. Precisión en la identificación bajo diferentes combinaciones de parámetros (C,γ) (Liu et al. 2019)

Los resultados indicaron que la mayor precisión de identificación lograda por el algoritmo propuesto fue del 98 %. Además, cuando se cumplen las condiciones mostradas en (1), el modelo SVM basado en el kernel de base radial proporciona un buen rendimiento de identificación de señal de fuga de tubería.

$$C \geq 2^2, \gamma \leq 2^0 \text{ y } 2^1 \leq C \times \gamma \leq 2^7 \qquad (1)$$

En (Alves Coelho et al. 2020) se analiza la eficiencia del árbol de decisión, cómo los datos que se ingresan en el algoritmo pueden afectar su precisión. Para esto, se crearon dos conjuntos dentro de los datos a analizar. Uno son los recopilados de acuerdo con la secuencia de sensores en el sistema, y el otro es donde se agruparon los datos para mostrar cómo funciona toda la ruta de las tuberías. Esto se hizo para comprender cuál de los conjuntos de datos presenta los mejores resultados para el escenario y el sistema previstos. Al crear el conjunto de datos, además de la marca de tiempo y los datos del sensor recopilados, se agregaron otras características a cada registro en función de los cálculos y los datos procesados previamente de los valores del sensor.

Para el escenario analizado anteriormente, se realizaron un total de 8 pruebas, cada una con diferentes parámetros, con el objetivo de entrenar el algoritmo y comprender bajo cuales circunstancias presenta la mejor precisión. La tabla 1 muestra todas las pruebas realizadas, identificando el conjunto de datos utilizado, el objetivo y las características no utilizadas

Tabla 1. Escenarios de prueba para obtener la efectividad del árbol de decisión en la detección de rupturas súbitas

Conjunto de datos	N° de Prueba	Características no usadas
Agrupados	1	Promedio de los últimos 5 valores para ese sensor, Diferencia con el sensor anterior, Diferencia del sensor 1
	2	Diferencia con el sensor anterior, Diferencia del sensor 1
	3	Diferencia del sensor 1
	4	_
Normal	5	Promedio de los últimos 5 valores para ese sensor, Diferencia con el sensor anterior, Diferencia del sensor 1
	6	Diferencia con el sensor anterior, Diferencia del sensor 1
	7	Diferencia del sensor 1
	8	_

La tabla 2 muestra los resultados de las pruebas realizadas donde la tendencia es que la precisión aumente con menos características no utilizadas

Tabla 2. Resultados de las pruebas de efectividad del árbol de decisión

Nro. de Prueba	1	2	3	4	5	6	7	8
Efectividad (%)	76,65	79,14	80,04	80,21	74,03	77,06	77,42	84,02

03 CONCLUSIONES

En los métodos basados en datos para la detección de rupturas súbitas en tuberías de agua no es necesario un conocimiento profundo del SDA, ya que estos solo implican análisis estadísticos o de procesamiento de señales de los datos adquiridos. Para SVM los resultados indicaron que la mayor precisión de identificación lograda por el algoritmo propuesto fue del 98 %. El árbol de decisión expuesto aumenta su precisión mientras más información del estado de la red tenga. Con la RNA se logra un error medio entre el 6 % y el 16 % del tamaño real, manteniendo el compromiso con el nivel de procesamiento.

04 REFERENCIAS

Alves Coelho J., Glória A. and Sebastião P. (2020). "Precise Water Leak Detection Using Machine Learning and Real-Time Sensor Data". IoT, 1(2), Article 2. ISSN 474-493.

Arredondo D. J., Gil W. J. y Mora J. J. (2017). "Metodología para la selección de atributos y condiciones operativas para la localización de fallas basada en la máquina de soporte vectorial". Tecnura, 21(51), 15-26. ISSN 0123-921.

Bohorquez J., Simpson A. R., Lamber, M. F. and Alexander B. (2021). "Merging Fluid Transient Waves and Artificial Neural Networks for Burst Detection and Identification in Pipelines". Journal of Water Resources Planning and Management, 147(1), 04020097. ISSN 1943-5452.

Hernández G. (2014). "Implementación numérica de redes neuronales artificiales para el análisis de grietas en placas". Tesis de Maestría, Escuela Superior de Ingeniería Mecánica y Eléctrica, México D.F.

Huang P., Zhu N., Hou D., Chen J., Xiao Y., Yu J., Zhang G. and Zhang, H. (2018). "Real-Time Burst Detection in District Metering Areas in Water Distribution System Based on Patterns of Water Demand with Supervised Learning". Water, 10(12), Article 12. ISSN 2073-4441.

Liu Y., Ma X., Li Y., Tie Y., Zhang Y. and Gao J. (2019). "Water Pipeline Leakage Detection Based on Machine Learning and Wireless Sensor Networks". Sensors, 19(23), Article 23. ISSN 1424-8220.

Mounce S. R., Mounce R. B. and Boxall J. B. (2011). "Novelty detection for time series data analysis in water distribution systems using support vector machines". Journal of hydroinformatics, 13(4), 672-686. ISSN 1464-7141.

Rodríguez C. F., Montes E. M. y López R. R. (2021). "Generación de árboles de decisión usando un algoritmo inspirado en la Física". Tesis de Maestría en Computación Aplicada, Laboratorio Nacional de Informática Avanzada, México D.F.

Saravanan R. and Sujatha P. (2018). A State of Art Techniques on Machine Learning Algorithms: A Perspective of Supervised Learning Approaches in Data Classification. 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), 945-949. ISBN 978-1-5386-2843-0.

Srirangarajan S., Allen M., Preis A., Iqbal M., Lim H. B. and Whittle A. J. (2013). "Wavelet-based Burst Event Detection and Localization in Water Distribution Systems". Journal of Signal Processing Systems, 72(1), 1-16. ISSN 1939-8115

Trutié-Carrero E., Valdés-Santiago D., León-Mecías Á. y Ramírez-Beltrán J. (2018). "Detección y Localización de Ruptura Súbita mediante Transformada Wavelet Discreta y Correlación Cruzada". Revista Iberoamericana de Automática e Informática Industrial, 15(2), Article 2. ISSN 1697-7912

Van Jaarsveldt C., Peters G. W., Ames M. and Chantler M. (2023). "Tutorial on Empirical Mode Decomposition: Basis Decomposition and Frequency Adaptive Graduation in Non-Stationary Time Series". IEEE Access. ISSN 2169-3536

Zaman D., Tiwari M. K., Gupta A. K. and Sen D. (2020). "A review of leakage detection strategies for pressurised pipeline in steady-state". Engineering Failure Analysis, 109, 104264. ISSN 1350-6307. https://doi.org/10.1016/j.engfailanal.2019.104264

CONFLICTO DE INTERESES

Los autores declaran que no existen conflictos de intereses.

CONTRIBUCIÓN DE LOS AUTORES

Jaime Ernesto Chiang Cruz https://orcid.org/0009-0001-4899-701X

Trabajó en el diseño de la investigación, búsqueda de información, comparación de algoritmos, interpretación de resultados, redacción y revisión del informe final

Iliover Vega González https://orcid.org/0000-0002-5811-994X

Participó en el diseño de la investigación, supervisión, experiencia, verificación de los resultados, revisión del informe final.

Jorge Ramírez Beltrán https://orcid.org/0000-0002-4125-2656

Participó en el diseño de la investigación, su conceptualización y metodología, supervisión, experiencia, verificación de los resultados, revisión del informe final.