PREDICCIÓN DE CAUDALES MEDIOS MENSUALES USANDO ÁRBOLES DE REGRESIÓN

Julián David Rojo Hernández, Luis Fernando Carvajal S y Oscar José Mesa S.

Departamento de Geociencias y Medio Ambiente, Facultad de Minas, Universidad Nacional de Colombia, Medellín – Colombia.
E-mail: jdrojoh@unal.edu.co, lfcarvaj@unal.edu.co, ojmesa@unal.edu.co

Resumen

El presente trabajo tiene por objeto introducir los árboles de decisión M5 para elaborar pronósticos informados de caudales medios mensuales mediante la incorporación de variables explicativas como el ENSO y las temperaturas superficiales del océano pacífico predichas por diferentes agencias.

Se introduce el concepto de árboles de decisión y se el esquema general del algoritmo M5 para la construcción de árboles de decisión basados en esquemas de regresión por inducción. Se explican los diferentes tipos de variables a ser utilizadas en la contracción de un árbol de decisión y se propone una metodológica para la incorporación de información climática mediante reglas de juicio usando M5. Dicha metodología es aplicada sobre el río Guadalupe en Colombia en diferentes horizontes de pronóstico para luego compararla con los resultados de otro modelo no lineal ampliamente utilizado como lo son las Redes Neuronales Artificiales. Los resultados indican que la predicción usando información macro-climática de diferentes fuentes dentro del esquema de árboles de regresión-decisión permite mejorar los indicadores de error de forma significativa para diferentes horizontes en comparación con las Redes Neuronales Artificiales.

Introducción

Un árbol de decisión es un modelo de predicción utilizado en el ámbito de la inteligencia artificial que sirve para codificar el conocimiento de un experto mediante la construcción de diagramas lógicos. Los árboles de decisión son sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva a fin de dar solución a un problema cuando se cuenta con buenos datos. Uno de los M5 fue propuesto de Quinlan (1992) para la solución de los árboles regresión-decisión está basado en el siguiente procedimiento: se divide el espacio dado por las variables de entrada en áreas (subespacios) y se construye sobre cada una de ellos un modelo especializado de regresión lineal múltiple. La construcción de los subespacios utiliza el concepto de árbol de decisión pero en lugar de una etiqueta de clasificación, este tendría una función de regresión, por lo que hacen un trabajo análogo al de los modelos de partición propuestos por Friedman como se muestra en la Figura 1.

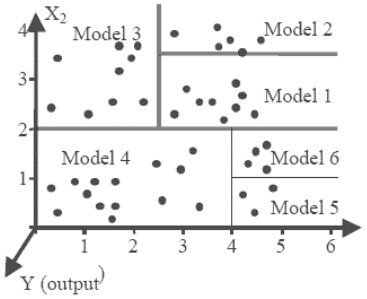

Figura 1.- Ejemplo de figura.

La aplicación del algoritmo M5 permite incorporar en los procesos de pronóstico tanto variables cuantitativas como cualitativas mediante reglas del tipo if – then. Finalmente la construcción de árboles de decisión para el pronóstico de caudales medios mensuales puede hacerse de forma empírica utilizando la experiencia que posee el modelador en el manejo de los diferentes esquemas de pronóstico además de la variabilidad de las predicciones en condiciones cambiantes de las variables explicativas.

Algoritmo de clasificación M5.

Suponga que se tiene una colección de *T* datos para calibrar un modelo de la forma:

$$y = f(X_1, X_2, ..., X_n) + \varepsilon \qquad [1]$$

Los modelos basados en árboles de decisión son construidos bajo el concepto de "divide y vencerás" mediante un criterio heurístico de agrupamiento (cluster) que busca minimizar la variación interna de los valores de la clase dentro de cada subconjunto Quinlan (1992), eligiendo aquel atributo que maximice la reducción de la variancia de cada subconjunto de acuerdo a la siguiente fórmula:

$$SDR = sd(T) - \sum_i \frac{|T_i|}{|T|} sd(T_i) \qquad [2]$$

Donde T corresponde a la colección de datos en el nodo a dividir, el conjunto de datos correspondiente al atributo i considerando en la división de T, y el operador $sd()$ estima la desviación típica de los datos. Finalmente la aplicación de mínimos cuadrados sobre cada una de las subregiones obtenidas permite hallar las relaciones existentes entre las diferentes variables involucradas en el problema.

Como datos de entrada a este tipo de modelo se pueden utilizar las variables cuantitativas, es decir, aquellas variables cuyos valores son un conjunto de cualidades no numéricas a las que se llama categorías o modalidades, entre ellas variables cualitativas ordinales las cuales permiten establecer relaciones de orden entre las categorías. (Ejemplo: Niño, Normal, Niña) o variables cualitativas por intervalos que permiten conocer la distancia numéricas entre dos niveles. (Ejemplo: Enero, Febrero, Marzo…) y además puede ser agrupada por intervalos (DEF, MAM,JJA,SON). También podrían utilizarse dentro de este modelo variables Binarias que solo adoptan valores de 0, 1 (si, no; existe, no existe, verdadero, falso) y se usan para resolver problemas del tipo inclusión-exclusión. Por ejemplo: llueve-no llueve; verano-invierno, y finalmente el modelo permite utilizar variables cuantitativas sean discretas o continuas.

Datos y metodología

Los datos utilizados para el desarrollo del presente trabajo corresponden a las serie de caudales medios mensuales del río Guadalupe, cuya estación de aforo se encuentra ubicada en inmediaciones de la población de Carolina del Príncipe al noroccidente de Colombia. El río Guadalupe constituye el eje central de la cadena de generación eléctrica conocida como GUATRON que actualmente genera el 7% de la hidroelectricidad del país y que pertenece a las Empresas Públicas de Medellín: el registro de la serie de caudales inicia en enero de

1938 y finaliza en diciembre de 2017 . La serie de caudales del río Guadalupe y su ubicación se presentan en la Figura 2.

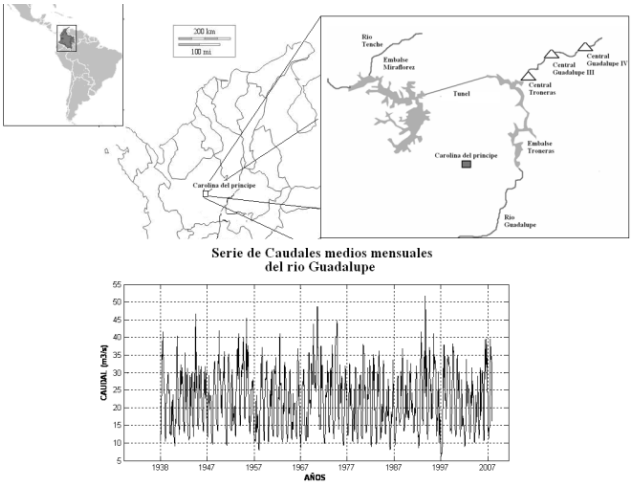

Figura 2.- Serie de Caudales medios mensuales del río Guadalupe.

El procedimiento para la predicción es el siguiente:

Identificar fuentes de información: La NOAA posee un centro de predicciones climáticas conocido como NCEP/NWS donde se resumen las predicciones de las temperaturas superficiales del océano para los diferentes modelos propuestos por varios autores desde 1980. Igualmente el Instituto de Clima y Sociedad de la Universidad de Columbia (IRI por sus siglas en inglés) ha desarrollado un modelo probabilístico estacional que estima la probabilidad de ocurrencia de los estados Niño, Normal, y Niña; la base de datos tiene un periodo de registro que comprende reportes desde el año 2003 a la fecha y los pronósticos están dados para 10 meses. Además los caudales poseen una importante componente estacional por lo que otra variable explicativa podría ser el mes específico de cada pronóstico.

Construcción de las variables del árbol de decisión: Para el presente trabajo se define como variable cualitativa los pronósticos del ENSO cuyos atributos son construidos en función de los valores de probabilidad asociados al pronóstico de cada una s de sus fases según el IRI (EL Niño=0; Neutral=1; La Niña=2). Los pronósticos de las SST serán incluidos como variable continua y los meses del año serán involucrados como variables cualitativas por intervalos como se muestra en la Figura 3.

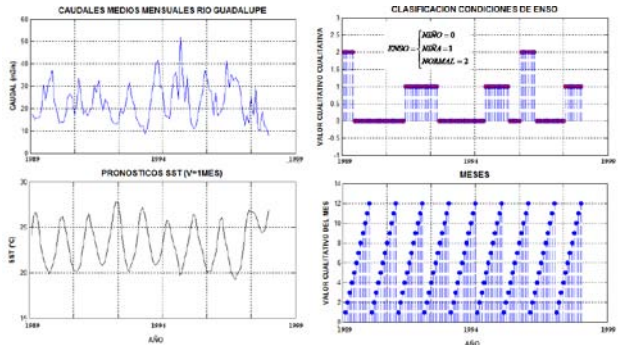

Figura 3.- Variables utilizadas en la construcción del árbol de decisión para el río Guadalupe.

Estrategia de calibración y validación: Con el ánimo de aplicar el algoritmo M5 a la modelación de los caudales del río Guadalupe se han definido como periodo de calibración la información entre el año 2002 y 2011 (10 años) y la validación entre 2012 y 2017 (5 años); y para analizar en mayor detalle se comparan los resultados obtenidos con aquellos generados por una Red Neuronal Artificial (RNA) del tipo perceptón multicapa, otro modelo no lineal entrenado para predecir caudales con las mismas variables de entrada.

Se creará una dispersión entre el valor predicho y el valor histórico, su cercanía a la recta de 45° permitirá analizar la existencia de errores sistemáticos en el pronóstico, el coeficiente de correlación entre ambos valores permitirá analizar cuanto se acercan los valores predichos a los valores históricos, además se estimará el valor medio porcentual de los errores (MAPE) y la raíz del error cuadrático medio (RMSE).

Resultados

Los resultados obtenidos luego de la aplicación del algoritmo M5 para el pronóstico de caudales medios mensuales a diferentes horizontes, muestran que el uso de árboles de decisión incorporando pronósticos variables explicativos permite disminuir significativamente el error de los pronósticos al incorporar de forma indirecta la información macro climática en comparación con otro modelo no lineal como las Redes Neuronales Artificiales, por ejemplo el indicador de error MAPE, con las RNA sería de 9.6% mientras que con el algoritmo M5 sería del 5.3% para el río Guadalupe como se muestra en la Figura 4 para un horizonte de pronóstico de un mes. Los errores asociados a otros horizontes son analizados en la Tabla 1.

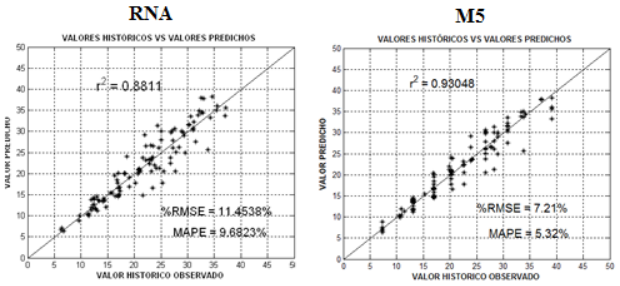

Figura 4.- Validación de los pronósticos usando Redes Neuronales Artificiales (RNA) y árboles de Decisión M5 para horizonte de un mes.

Tabla 1.- Comparación de los errores de pronóstico entre árboles de regresión (M5) y Redes Neuronales artificiales (RNA).

Modelo	RNA		M5	
Horizonte	RMSE(%)	MAPE(%)	RMSE(%)	MAPE(%)
1	11.45	9.68	7.21	5.32
3	13.73	11.73	9.11	7.96
6	15.27	14.72	11.31	11.32

Conclusiones

Los resultados anteriores demuestran que el algoritmo M5 representa adecuadamente los caudales del río Guadalupe gracias a su capacidad para generar regresiones locales mediante la partición del espacio formado por las variables explicativas. Dicho enfoque presenta ventajas significativas cuando se compara con otros métodos no lineales como las Redes Neuronales Artificiales, siendo en este caso, lo errores derivados de los árboles de regresión menores y bastante adecuados para la predicción de caudales del río Guadalupe.

Referencias

J.R. Quinlan, (1992). Learning withcontinuous classes, *Proceedings of the Australian Joint Conference on Artificial Intelligence, World Scientific, Singapore, pp. 343–348.*