Una modificación de la metodología de regresión simbólica para la predicción de series de tiempo[1]

A modification of the methodology of symbolic regression for time series prediction[2]

Alteração da metodologia de regressão simbólica para a predição de séries de tempo[3]

Carlos A. Martínez[4]
Juan D. Velásquez-Henao[5]

SICI: 0123-2126(201307)17:2<325:MMRSPS>2.0.TX;2-C

[1] Fecha de recepción: 8 de noviembre de 2011. Fecha de aceptación: 4 de abril de 2013. Este artículo se deriva de la tesis de maestría *Problemas abiertos en la aplicación de la regresión simbólica en el pronóstico de series de tiempo*, elaborada por C. A. Martínez y dirigida por J. D. Velásquez. Desarrollado por el Grupo de Investigación de Sistemas de la Universidad Nacional de Colombia, Medellín, Colombia.

[2] Received: November 8, 2011. Accepted: April 4, 2013. This article is derived from the master's thesis entitled "*Problemas abiertos en la aplicación de la regresión simbólica en el pronóstico de series de tiempo*" (Open problems in the application of symbolic regression in time series forecasting), developed by C. A. Martínez and directed by J. D. Velásquez. Developed by the the Systems Research Group at the Universidad Nacional of Colombia, Medellín, Colombia.

[3] Data de recebimento: 8 de novembro de 2011. Data de aceitação: 4 de abril de 2013. Este artigo é derivado da tese de mestrado *Problemas abertos na aplicação da regressão simbólica no prognóstico de séries de tempo*, elaborada por C. A. Martínez e dirigida por J. D. Velásquez. Desenvolvido pelo Grupo de Pesquisa de Sistemas da Universidade Nacional da Colômbia, Medellín, Colômbia.

[4] Ingeniero de sistemas e informática, magíster en Ingeniería de Sistemas y estudiante de doctorado de la Universidad Nacional de Colombia Medellín, Colombia. Correo electrónico: amartin@unal.edu.co.

[5] Ingeniero civil, magíster en Ingeniería de Sistemas y doctor en Ingeniería - Sistemas Energéticos, de la Universidad Nacional de Colombia. Profesor titular del Departamento de Ciencias de la Computación y la Decisión, Facultad de Minas, Universidad Nacional de Colombia, Medellín, Colombia. Correo electrónico: jdvelasq@unal.edu.co.

Resumen

En este artículo se propone una nueva metodología para la predicción de series de tiempo no lineales usando programación genética. La aproximación propuesta se basa en la incorporación del concepto de bloques funcionales y la modificación del algoritmo genético para que opere con estos. Los bloques funcionales representan modelos estadísticos bien conocidos para el pronóstico de series de tiempo. El algoritmo propuesto permite la exploración y explotación de regiones donde hay mayor posibilidad de encontrar mejores modelos de pronóstico. Para validar la aproximación propuesta, se pronosticaron dos series de tiempo Benchmark; se encontró que nuestra metodología pronostica con mayor precisión las series de tiempo consideradas en comparación con otros modelos no lineales.

Palabras clave

Pronóstico, programación genética, redes neuronales artificiales, algoritmos genéticos, modelos no lineales.

Abstract

In this paper we propose a new methodology for the prediction of nonlinear time series using genetic programming. The proposed approach is based on incorporating the concept of functional blocks and the modification of the genetic algorithm so that it operates with it. The functional blocks represent well known statistical models for the time series forecasting. The proposed algorithm allows the exploration and exploitation of regions where there is greater possibility of finding better forecasting models. Two Benchmark time series were predicted in order to validate the proposed approach, and it was found that our methodology predicts more accurately the time series considered, in comparison with other nonlinear models.

Keywords

Forecasting, genetic programming, artificial neural networks, genetic algorithms, nonlinear models.

Resumo

Neste artigo propõe-se nova metodologia para predição de séries de tempo não-lineares utilizando programação genética. A abordagem proposta baseia-se na incorporação do conceito de blocos funcionais e a alteração do algoritmo genético para que opere com estes. Os blocos funcionais representam modelos estadísticos bem conhecidos para o prognóstico de séries de tempo. O algoritmo proposto permite a especulação e exploração de regiões onde há maior possibilidade de encontrar melhores modelos de prognóstico. Para avaliar a aproximação proposta, prognosticaram-se duas séries de tempo Benchmark; encontro-se que nossa metodologia prognostica com maior precisão as séries de tempo consideradas em comparação com outros modelos não-lineares.

Palavras-chave

Prognóstico, programação genética, redes neuronais artificiais, algoritmos genéticos, modelos não-lineares.

1. Introducción

Aunque la regresión simbólica (SR, por sus siglas del inglés) fue propuesta originalmente como una técnica para aproximar de forma empírica la relación no lineal, existente y desconocida entre un conjunto de variables de entrada y una de salida (Koza, 1992), recientemente ha sido aplicada con éxito al pronóstico de series de tiempo. Por ejemplo, Wang, Chau, Cheng y Qiu (2009) comparan la PG con varias técnicas de inteligencia artificial en el pronóstico de series de tiempo hidrológicas y concluyen que la PG podría superar dichas técnicas alternativas en términos de su precisión; Abdelmalek, Hamida y Abid (2009) usan la PG para obtener modelos de pronóstico de la volatilidad del índice S&P500.

La PG también ha sido usada en el desarrollo de modelos híbridos de pronóstico: Chen, Pang, Wang y Xu (2008) combinan la PG y un modelo de corrección de errores vectoriales; Lee y Tong (2011) usan un modelo ARIMA para extraer la componente lineal de la serie de tiempo estudiada, y luego la PG es utilizada para extraer la componente no lineal remanente de los residuales del modelo ARIMA. Adicionalmente, la PG también ha sido utilizada para determinar la estructura óptima de modelos de redes neuronales artificiales (Nikolaev e Iba, 2003; De Menezes y Nikolaev, 2006; Bernal-Urbina y Flores-Méndez, 2008). Finalmente, Wagner, Khouja, Michalewicz y McGregor (2007, 2008) desarrollan un nuevo modelo de PG para ambientes dinámicos.

Las razones de su éxito se derivan de la capacidad para encontrar relaciones no lineales ocultas en los datos, las cuales no podrían ser adecuadamente representadas por técnicas paramétricas convencionales, sin la necesidad de suponer, a priori, la forma matemática de la no linealidad; en su capacidad para descartar las variables irrelevantes durante la ejecución del algoritmo; y en la facilidad para controlar problemas comunes en las técnicas no paramétricas, como el *data-snooping*, el cual consiste en alcanzar resultados superiores, pero que son obtenidos únicamente por aleatoridad y que ocurre cuando un mismo conjunto de datos es usado más de una vez para inferencia o la selección del modelo final (Sullivan *et al.*, 1999).

En SR, este problema puede controlarse fácilmente: por una parte, el algoritmo permite controlar la complejidad de las funciones encontradas al imponer un límite máximo a la profundidad de los árboles sintácticos que las representan; y segundo, al dividirse la información existente en datos para entrenamiento y pronóstico, aquellos modelos que presenten sobreajuste (un error de ajuste muy bajo para la muestra de entrenamiento) presentarán un error de pronóstico muy alto.

Una desventaja de la SR es que no permite garantizar que el algoritmo genético encuentre funciones con un desempeño mejor que los modelos tradicionales comúnmente usados y aquellos de uso difundido pronóstico de series de tiempo, como los modelos ARIMA o los basados en regímenes, aunque efectivamente exista una dinámica no lineal en la serie de tiempo. Dicha desventaja puede ocurrir porque los modelos matemáticos bien conocidos y comúnmente usados en econometría y estadística no se consideran explícitamente en el espacio de búsqueda; y porque los modelos propios de la literatura de series de tiempo pueden tener una estructura funcional compleja cuando son representados como árboles sintácticos, y resulta difícil y altamente improbable (en el sentido matemático) que el algoritmo genético genere aleatoriamente individuos con dicha estructura o que ella sea generada espontáneamente durante la corrida del algoritmo. En consecuencia, no es posible aprovechar el algoritmo genético como un mecanismo que busque heurísticamente modelos empíricos a partir de modelos ya conocidos en la literatura de series de tiempo, ya que concentra la exploración sobre regiones que contienen modelos de desempeño inferior a los modelos clásicos de la literatura de series de tiempo.

El primer objetivo de este trabajo es presentar el concepto de *bloque funcional* (BF), que en este trabajo se define como la unidad funcional básica a partir de la cual se pueden construir modelos de series de tiempo; estos BF son usados para la representación de los individuos en el algoritmo de PG. El segundo objetivo es modificar los operadores genéticos tradicionales usados en PG para que el algoritmo genético pueda operar sobre los bloques funcionales, y así, se optimice el proceso de búsqueda de la función empírica que mejor aproxime la dinámica de la serie de tiempo.

El resto del artículo está organizado como sigue: en la sección 2 se introducen las modificaciones al algoritmo original de Koza (1992), a partir del uso de BF y la modificación a los operadores de selección y cruce; en la sección 3 se realiza una verificación de los cambios propuestos con dos series Benchmark; y, finalmente, en la sección 4 se muestran las conclusiones y el trabajo futuro.

2. Metodología propuesta

La base conceptual de la RS parte de la aplicación de los operadores genéticos de reproducción, cruce y mutación a una población de individuos que codifican funciones matemáticas mediante sus árboles sintácticos equivalentes, cuyo espacio de búsqueda corresponde a todas las posibles funciones que son obtenibles a partir del conjunto de funciones definidas para los nodos intermedios, y las variables y constantes numéricas definidas para los nodos terminales. La metodología busca encontrar aquella función que aproxima con la mayor precisión posible una variable de salida cuando los valores de los parámetros de la función son conocidos. En dichos árboles sintácticos, los nodos internos son llamados nodos funcionales, mientras que los nodos en el final de las ramas son llamados nodos terminales. Cada nodo terminal es definido como una variable de entrada o una constante numérica (también llamada parámetro). Cada nodo interior corresponde a una función elegida de un conjunto de funciones primitivas o básicas predefinidas, como la suma, la resta, la multiplicación o la división.

En esta sección se describe una modificación de la metodología de regresión simbólica que tiene como finalidad incorporar las condiciones particulares del problema de predicción de series de tiempo no lineales.

2.1. Definición de los individuos

La metodología propuesta está basada en el concepto de *bloques funcionales* (BF), pues el conjunto de terminales está conformado únicamente por diferentes tipos de BF. Los BF son los constituyentes básicos de las ecuaciones; corresponden a las funciones más básicas que se pueden manejar como terminales de un individuo. De esta forma, pueden ser evaluados numéricamente sin depender de otras funciones externas. Todas las funciones definidas para los nodos interiores del árbol sintáctico pueden operar directamente sobre los BF sin modificaciones, ya que estos ocupan únicamente los nodos terminales.

En su forma más simple, los BF pueden ser definidos como una función $B(\cdot)$ que corresponde a la combinación lineal de las funciones $g_i(\cdot)$:

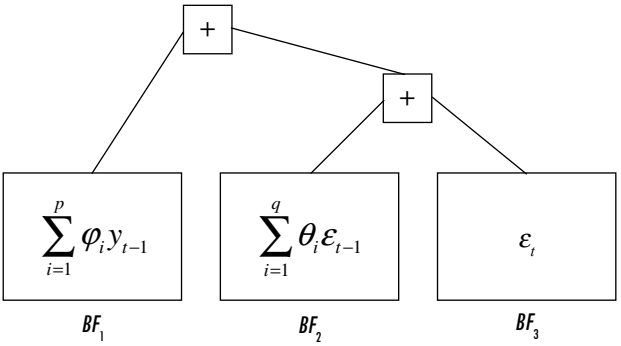
$$BF = b + \sum_i w_i g_i\left(X\right) \tag{1}$$

Donde X es el vector de entradas (variables/rezagos) del modelo, y b y w_i son los parámetros del BF. Es posible definir BF más complejos a partir de la combinación de BF más simples. A continuación se ejemplifica el uso de los BF. Al ser, por ejemplo, el modelo ARMA (p, q) definido como:

$$y_t = \sum_{i=1}^{p} \varphi_i y_{t-1} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t = BF_1 + BF_2 + BF_3 \quad (2)$$

Si el BF_1 es definido como la parte autorregresiva (AR) (con $i = \{1,..,p\}$, $b = 0$, $w_i = \varphi_i$ y $g_i(X) = y_{t-i}$), el BF_2 (con $i = \{1,..,q\}$, $b = 0$, $w_i = \theta_i$ y $g_i(X) = e_{t-i}$) como la componente de promedios móviles (MA), y el BF_3 (con $i = \{1\}$, $b = 0$, $w_i = 1$ y $g_i(X) = E_t$), como la componente aleatoria, entonces el modelo anterior puede ser representado como la suma de tres BF y equivale al árbol sintáctico presentado en la figura 1.

Figura 1. Representación en forma de árbol con bloques funcionales de un modelo ARMA



Fuente: presentación propia de los autores.

Entre las principales características de los BF se encuentran las siguientes: primero, permiten expresar ecuaciones complejas de forma simple, como ya se ilustró; ello facilita la aplicación de los operadores genéticos, evita la redundancia de operadores (en la versión clásica de GP serían necesarios $2 * (p + q) + 1$ nodos, mientras en la versión de BF solo son necesarios cinco nodos) y permite una exploración más amplia del campo de búsqueda. Segundo, los individuos resultantes pueden interpretarse como modelos híbridos que combinan de forma novedosa los modelos básicos definidos por los BF. Tercero, los individuos pueden interpretarse en términos de los BF que los conforman, por consiguiente, en términos de los modelos bien establecidos en la literatura, con lo cual se gana en claridad.

2.2. Algoritmo propuesto

El algoritmo de SR usado en esta investigación está conformado por los siguientes pasos:

1. Inicializar los parámetros del modelo, entre los que se cuentan los operadores, tipos de BF, iteraciones, criterios de parada y funciones de optimización que van a ser utilizadas.
2. Se genera la población inicial $P = \{S_1,\ldots, S_\mu\}$ de μ individuos utilizando los BF definidos como funciones base de los terminales. Dichos individuos son generados aleatoriamente de acuerdo con un vector de terminales T, un vector de operadores F y un número máximo de nodos y niveles (profundidad del árbol definida en el paso anterior).
3. Se evalúa la función de aptitud $f(S_i)$ para cada uno de los individuos S_i, $i = 1,\ldots, \mu$, de la población actual P.
4. Se genera la nueva población P^* de μ hijos, aplicando los operadores genéticos de elitismo, clonación, cruce y mutación a la población actual P.
5. Se reemplaza la población actual por la nueva población generada, de acuerdo con el criterio de selección definido.
6. Se evalúan los criterios de parada, si no se cumplen, se vuelve al paso 3. En caso contrario, se continúa con el paso 7.
7. Se evalúa la función de intensificación, la cual consiste en aplicar un algoritmo de optimización al mejor individuo encontrado en población actual, tomando como punto inicial aquel compuesto por los parámetros encontrados para este. Ello permite una mejora en las medidas de error de aproximación, manteniendo la estructura de la solución.

2.3. Operadores genéticos

- Función por minimizar: en vez de minimizar el error cuadrático medio, como se realiza tradicionalmente, se minimizó el criterio de información de Akaike, con el fin de castigar la complejidad de la expresión resultante y obligar al algoritmo a buscar modelos parsimoniosos.
- Esquema de selección de padres: probabilístico proporcional a la aptitud.
- Función de aptitud: torneo estocástico.
- Elitismo: los mejores κ individuos son copiados a la población de hijos sin mutar.
- Cruce: la mitad de la población de hijos es obtenida usando el operador definido por Koza (1992). Para cada pareja de padres, se selecciona aleatoriamente un nodo de cruce en cada padre y se intercambian los subárboles correspondientes. Para la otra mitad de la población de hijos se consideró un nuevo operador de cruce, en el que se obtiene un único hijo a partir de la suma algebraica de dos padres.

- Mutación: se implementó el operador de mutación propuesto por Sette y Boullart (2001), en el que se selecciona aleatoriamente un nodo del árbol y se genera aleatoriamente un nuevo subárbol.
- Simplificación de los árboles sintácticos: se implementaron las siguientes reglas de simplificación, que difieren de las reglas propuestas originalmente por Koza (1992), con el fin de facilitar la exploración de funciones matemáticas: $a * x \pm b * x = c * x$, $a * x/b * x = c * x$, $a * b * x = c * x$, donde a, b, c son constantes y x es el terminal correspondiente.
- Estimación de los parámetros óptimos de cada individuo: a diferencia del trabajo original de Koza (1992), en este manuscrito se considera únicamente la forma funcional de los individuos, de forma que para cada individuo generado durante el proceso se hace la estimación de sus parámetros óptimos; esto es realizado mediante la OPTIM implementada en el lenguaje R. Ello causa que si hay dos individuos con la misma forma funcional en la población actual, el algoritmo de optimización numérica genera los mismos valores para sus parámetros; por consiguiente, ambos individuos tienen el mismo valor de la función de aptitud. Igualmente, cada cierto número de generaciones se aplica un algoritmo numérico de búsqueda global para intentar mejorar los parámetros de cada individuo de la población, con el fin de garantizar una exploración adecuada del espacio de valores de los parámetros. En este caso se usó la subrutina RGNOUD del lenguaje R.

3. Experimentos numéricos

3.1. Serie AIRLINE

Esta serie, también conocida como serie *G* de Box y Jenkins (1970), corresponde al logaritmo natural del número de pasajeros transportados mensualmente al exterior por una aerolínea entre ENE1949 y DIC1960. La serie tiene 132 observaciones; las primeras 120 son usadas para la estimación de los modelos de pronóstico y las 12 restantes, para la evaluación de su capacidad predictiva.

La serie ha sido pronosticada en diferentes estudios. Faraway y Chatfield (1998) compararon un modelo SARIMA $(0,1,1)(0,1,1)_{12}$ y diferentes configuraciones de redes neuronales tipo perceptrón multicapa (MLP), que difieren en los rezagos considerados y en el número de neuronas en la capa oculta; la sumatoria del error cuadrático (SSE, por sus siglas del inglés) para las muestras de entrenamiento y predicción reportadas por Faraway y Chatfield (1998) es reportada en la tabla 1. Velásquez, Olaya y Franco (2010) pronosticaron esta serie usando

máquinas de vectores de soporte (SVM, por sus siglas del inglés); en la tabla 1 se reproducen las estadísticas informadas por Velásquez *et al.* (2010) para el SVM con menor SSE para la muestra de entrenamiento y el SVM con menor SSE para la muestra de predicción; la motivación para reportar los resultados de estos dos modelos es que dichas magnitudes permiten establecer una cota empírica para el SSE que permitiría juzgar el desempeño de otros modelos.

En nuestra investigación, esta serie fue pronosticada usando el algoritmo original de PG (Koza, 1992); la metodología desarrollada en este trabajo, usando los primeros 120 datos para la estimación del modelo, y los 12 restantes para evaluar su capacidad predictiva.

Tanto la versión original del algoritmo de regresión simbólica como el propuesto en este artículo fueron ejecutados varias veces con el fin de encontrar la mejor combinación de parámetros para cada serie de tiempo. Para el caso de la serie AIRLINE, se encontró que el modelo con menor SSE para la muestra de entrenamiento (o calibración) es obtenido usando los siguientes parámetros:

- Población inicial: 20 individuos.
- Nivel inicial de profundidad de los individuos: 3 con 7 terminales máximo.
- Máximo número de generaciones: 10.
- Función de error: SSE.
- Algoritmo de optimización: OPTIM.
- Algoritmo de profundización RGNOUD (implementación de algoritmos genéticos).

Tabla 1. Resultados de entrenamiento y pronóstico entre modelos de predicción para la serie AIRLINE

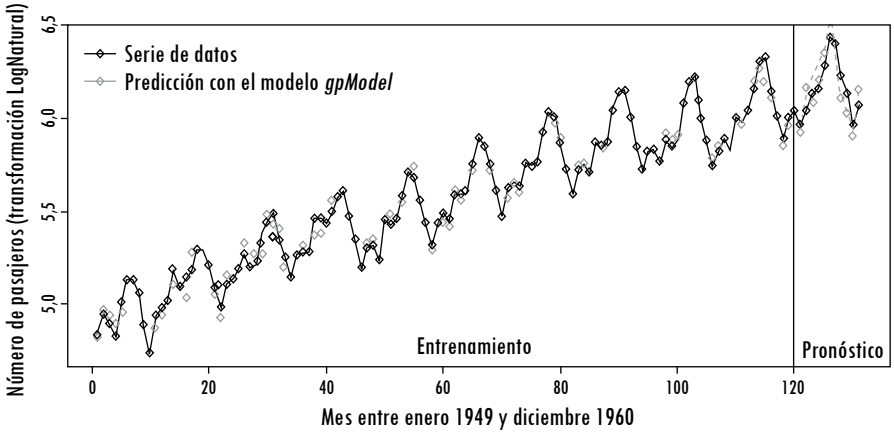
Modelo	Rezagos	SSE Entrenamiento	SSE Predicción
SARIMA (Faraway y Chatfield, 1998)	1, 12, 13	1,08	0,43
MLP con menor SSE en entrenamiento (Faraway y Chatfield, 1998)	1-13	0,26	1,12
MLP con menor SSE en pronóstico (Faraway y Chatfield, 1998)	1, 12	2,30	0,34
SVM con menor SSE en entrenamiento (Velásquez *et al.*, 2010)	1, 2, 12, 13	0,18	0,20
SVM con menor SSE en pronóstico (Velásquez *et al.*, 2010)	1-13	0,20	0,00
GP tradicional (este estudio)	1-13	1,32	0,80
GP propuesta (este estudio)	1-13	0,17	0,01

Fuentes: Faraway y Chatfield (1998), Velásquez *et al.* (2010) y cálculos de los autores.

En la tabla 1 se reportan los resultados obtenidos en los cuales se aprecia que con la metodología propuesta se obtiene el menor SSE de entrenamiento, y se resalta una mejora del 84,3% respecto al modelo ARIMA. Por otro lado, al analizar los resultados del SSE de predicción se puede notar que nuestra metodología solo es superada por la SVM con menor SSE para la muestra de pronóstico (diferencia de 0,01 en el SSE registrado); nótese que este SVM no presenta el mejor ajuste a la muestra de entrenamiento. Aunque en comparación a las demás técnicas hay una disminución de la medida SSE del 1,900% (SVM con menor SSE en entrenamiento) al 11,100% (MLP con menor SSE en entrenamiento).

En este caso, se concluye que el algoritmo de PG propuesto en este artículo representa mejor la dinámica de la serie, ya que presenta el mejor ajuste a la muestra de entrenamiento y una precisión similar en el pronóstico respecto a otros modelos reportados en la literatura. En la figura 2 se presenta la serie original y la predicción de la metodología propuesta para las muestras de entrenamiento y predicción.

Figura 2. Resultados de entrenamiento y pronóstico del modelo de predicción para la serie AIRLINE por medio de la metodología propuesta



Fuente: presentación propia de los autores.

3.2. Serie LYNX

La serie LYNX corresponde al número de linces canadienses atrapados por año en el distrito del río Mackenzie del norte de Canadá, entre 1821 y 1934 (114 observaciones). Para su pronóstico, la serie es transformada con el logaritmo base 10; los primeros 100 datos para el entrenamiento y los 14 restantes para la predicción.

Esta serie fue utilizada por Zhang (2003) para comparar la capacidad de predicción de un modelo ARIMA, una red neuronal artificial (con siete neuronas de entrada y cinco neuronas en la capa oculta) y un modelo híbrido que combina las dos técnicas anteriores. El MSE y la desviación media absoluta (MAD) para las muestras de entrenamiento y predicción calculadas por Zhang (2003) son reportados en la tabla 2. Esta serie también fue utilizada por Velásquez *et al.* (2010) para evaluar el desempeño de las SVM. En la tabla 2 se reproducen los resultados de Velásquez *et al.* (2010) para las SVM con menores MSE de entrenamiento y validación, respectivamente.

En nuestra investigación, esta serie fue pronosticada usando el algoritmo de PG original y la modificación propuesta. Los parámetros óptimos para la corrida (que generan el menor SSE para la muestra de entrenamiento) fueron los siguientes:

- Población inicial: 20 individuos.
- Nivel inicial de profundidad de los individuos: 3 con 7 terminales máximo.
- Máximo número de generaciones: 10.
- Función de error: MSE.
- Algoritmo de optimización: OPTIM.
- Algoritmo de profundización: RGNOUD.

Tabla 2. Resultados de entrenamiento y pronóstico entre modelos de predicción para la serie LYNX

Modelo	Rezagos	MSE (MAD) Entrenamiento	MSE (MAD) Predicción
ARIMA (Zhang, 2003)	N/D	N/D	0,021 (0,112)
MLP (Zhang, 2003)	N/D	N/D	0,021 (0,112)
Híbrido (Zhang, 2003)	N/D	N/D	0,017 (0,104)
SVM con menor MSE en entrenamiento (Velásquez *et al.*, 2010)	1-9	0,026 (0,140)	0,036 (1,163)
SVM con menor MSE en pronóstico (Velásquez *et al.*, 2010)	1-3, 8-10	0,034 (0,152)	0,015 (0,087)
GP tradicional (este estudio)	1-10	6,810 (19,240)	1,980 (3,150)
GP propuesta (este estudio)	1-10	0,028 (0,115)	0,017 (0,102)

Fuentes: Zhang (2003), Velásquez *et al.* (2010) y cálculos de los autores.

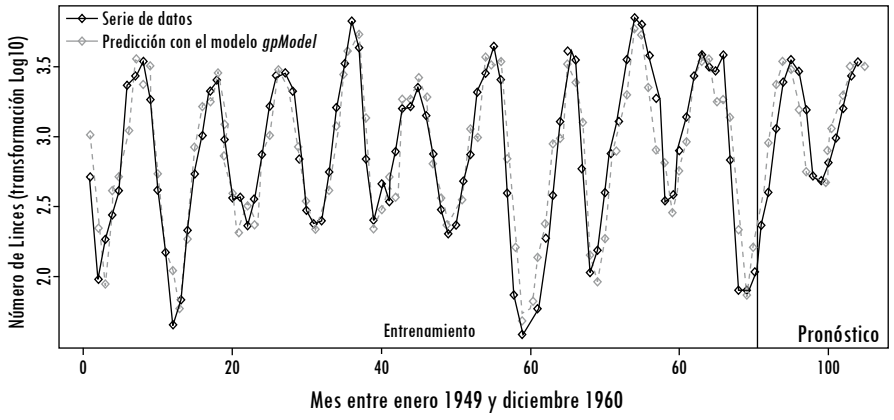
Las estadísticas de error obtenidas son informadas en la tabla 2. Se aprecia que la metodología propuesta solo es superada en precisión para la muestra de entrenamiento por el SVM presentado por Velásquez *et al.* (2010), pero dicho modelo (el SVM) tiene muy poco poder predictivo, lo cual es un indicativo de

un posible sobreajuste a los datos de entrenamiento. Para los modelos ARIMA, MLP y el modelo híbrido de Zhang (2003) no registran datos sobre su precisión de ajuste a la muestra de entrenamiento.

Por otra parte, la mayor precisión para la muestra de pronóstico es alcanzada por el otro SVM presentado por Velásquez *et al.* (2010), pero este modelo tiene un ajuste muy bajo a la muestra de calibración, por lo cual el SSE de pronóstico puede usarse únicamente como un indicativo empírico de la máxima precisión que podría alcanzarse por un modelo. En consecuencia, puede decirse que con la metodología propuesta se alcanza una precisión similar a los mejores modelos informados en la literatura para esta serie de tiempo.

Al analizar los datos de predicción, la técnica propuesta posee un desempeño similar al modelo híbrido de Zhang (2003); al analizar los demás resultados de predicción, se resaltan las mejoras respecto a los modelos ARIMA y MLP, y sobre el algoritmo tradicional de SR. Se concluye, entonces, que la metodología propuesta supera las demás técnicas informadas en la literatura, ya que su precisión sobre la totalidad de la serie de datos es superior. En la figura 3 se presenta la serie original y la predicción de la metodología propuesta para las muestras de entrenamiento y predicción.

Figura 3. Resultados de entrenamiento y pronóstico del modelo de predicción para la serie LYNX por medio de la metodología propuesta



Fuente: presentación propia de los autores.

4. Conclusiones y trabajo futuro

Se introdujo el concepto de bloque funcional para la representación de modelos de series de tiempo y se modificó la metodología tradicional de regresión sim-

bólica para adaptarla al uso de bloques funcionales. Los experimentos numéricos desarrollados permiten concluir, al menos para los casos Benchmark considerados, que las modificaciones propuestas impactan positivamente el desempeño del algoritmo, de tal manera que es posible encontrar empíricamente nuevos modelos matemáticos que pueden pronosticar con mayor precisión las series estudiadas, en comparación con otros modelos previamente usados en la literatura y el algoritmo tradicional de programación genética.

Aunque los resultados obtenidos son prometedores, la metodología dista mucho de estar finalizada. Por ello, como trabajo futuro se plantea la necesidad de trabajar en diferentes aspectos, como la eficiencia computacional de la implementación, la mejora de los algoritmos de optimización para calcular los parámetros de los modelos, la incorporación de nuevos operadores genéticos y la consideración de cambios estructurales en los datos.

Referencias

ABDELMALEK, W.; HAMIDA, S.B. y ABID, F. Selecting the best forecasting-implied volatility model using genetic programming. *Journal of Applied Mathematics and Decision Sciences*. 2009, vol. 2009. art. no. 179230.

BERNAL-URBINA, M. y FLORES-MÉNDEZ, A. Time series forecasting through polynomial artificial neural networks and genetic programming. *Proceedings of the International Joint Conference on Neural Networks*. 2008, art. no. 4634270, pp. 3325-3330

BOX, G. E. P. y JENKINS, G. M. *Time Series Analysis: Forecasting and Control*. 3rd Ed. Englewood Clifs, NJ: Prentice Hall, 1970.

CHEN, X.; PANG, Y.; WANG, S.Y., et al. A new integrated forecasting method. *Xitong Gongcheng Lilun yu Shijian/System Engineering Theory and Practice*. 2008, vol. 28, núm. 4, pp. 108-112+123.

DE MENEZES, L. M. y NIKOLAEV, N. Y. Forecasting with genetically programmed polynomial neural networks. *International Journal of Forecasting*. 2006, vol. 22, núm. 2, pp. 249-265.

FARAWAY, J. y CHATFIELD, C. Time series forecasting with neural networks: a comparative study using the airline data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. 1998, vol. 47, pp. 231-250.

KOZA, J. R. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge: MIT Press, 1992.

LEE, Y. S. y TONG, L. I. Forecasting time series using a methodology based on autoregressive integrated moving average and genetic programming. *Knowledge-Based Systems*. 2011, vol. 24, núm. 1, pp. 66-72.

NIKOLAEV, N. Y. y IBA, H. Learning polynomial feedforward neural networks by genetic programming and backpropagation. *IEEE Transactions on Neural Networks*. 2003, vol. 14, núm. 2, pp. 337-350.

SETTE, S. y BOULLART, L. Genetic Programming: Principles and applications. *Engineering Applications of Artificial Intelligence*. 2001, vol. 14, núm. 6, pp.727-736.

SULLIVAN, R.; TIMMERMAN, A. G y WHITE, H. L. Jr. Data-Snooping, Technical Trading Rule Performance, and the Bootstrap. *Journal of Finance*. 1999, vol. 54, núm. 5, pp. 1647-1691.

VELÁSQUEZ, J. D.; OLAYA, Y.; FRANCO, C. J. Predicción de series temporales usando Máquinas de Vectores de Soporte. *Ingeniare, Revista Chilena de Ingeniería*. 2010, vol. 18, núm. 1, pp. 64-75.

WANG, W. C.; CHAU, K. W.; CHENG, C. T., et al. A comparison of performance of several artificial intelligence methods for forecasting monthly discharge time series. *Journal of Hydrology*. 2009, vol. 374, núm. 3-4, pp. 294-306.

WAGNER, N.; MICHALEWICZ, Z.; KHOUJA, M., et al. Time series forecasting for dynamic environments: The DyFor genetic program model. *IEEE Transactions on Evolutionary Computation*. 2007, vol. 11, núm. 4, pp. 433-452.

WAGNER, N.; KHOUJA, M.; MICHALEWICZ, Z., et al. Forecasting economic time series with the DyFor genetic program model. *Applied Financial Economics*. 2008, vol. 18, núm. 5, pp. 357-378.

ZHANG, G. Time Series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*. 2003, vol. 50, pp. 159-175.