Pronóstico de demanda por medio de redes neuronales artificiales

María Angélica Salazar Aguilar, Mauricio Cabrera Ríos
División de Posgrado en Ingeniería de Sistemas, FIME-UANL
mcabrera@uanl.mx

RESUMEN

En este trabajo se describe la utilización de Redes Neuronales Artificiales (RNAs) para pronóstico de demanda. Se propone además un método para definir los parámetros de las RNAs de una manera integrada y repetible y se prueba con una aplicación real.

PALABRAS CLAVE

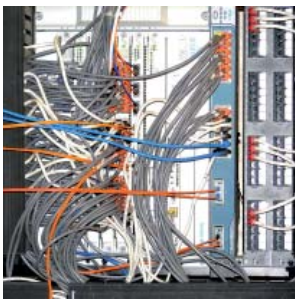
Pronóstico de demanda, Redes Neuronales Artificiales, RNAs, optimización.

ABSTRACT

The use of Artificial Neural Networks (ANNs) for demand forecasting is described in this work. A novel method to define the ANNs parameters in an integrated and repeatable fashion is proposed and demonstrated through a case study in a local company.

KEYWORDS

Demand forecasting, Artificial Neural Networks, ANNs, optimization.



INTRODUCCIÓN

En toda industria, la planeación es una necesidad. Un objetivo importante de la planeación es tratar de prever lo que puede suceder en el futuro. En este trabajo, se colaboró con una empresa de telecomunicaciones con necesidad de planear a nivel operacional, estratégico y táctico para mantenerse competitiva ante las fluctuaciones de mercado y cursos de acción de sus competidores. Esta empresa, como la gran mayoría, tiene como objetivo principal generar utilidades y brindar un alto nivel de servicio a sus clientes.

El principal recurso de la empresa es una red de transmisión con capacidad finita. Los clientes demandan la utilización de esta red en forma estocástica.

Para cumplir con el alto nivel de servicio y maximizar las utilidades, la red de transmisión debe tener capacidad suficiente para satisfacer la demanda de los clientes. Por ello, le corresponde al tomador de decisiones determinar la capacidad de la red.

A partir de un pronóstico, el tomador de decisiones puede determinar la capacidad que se requiere en la red de transmisión para satisfacer la demanda, así como determinar con anticipación si es necesaria una expansión de capacidad. Un buen trabajo de pronóstico deberá resultar en una mejor planeación del presupuesto anual, así como un mejor aprovechamiento de los recursos económicos de la empresa.

Artículo basado en la tesis "Pronóstico de demanda por medio de redes neuronales artificiales (RNAs) en la industria de telecomunicaciones" galardonada con el Premio a Mejor Tesis de Maestría UANL 2006 en la categoría de Ingeniería y Tecnología.

Para hacer un pronóstico es común requerir información cuantitativa del comportamiento de la demanda a través del tiempo, es decir, una serie de tiempo, siendo el Análisis de Series de Tiempo la técnica estadística más utilizada para estimar su comportamiento.

Por muchos años, este tipo de análisis ha estado dominado por la utilización de métodos estadísticos lineales que se pueden implementar de manera conveniente, sin embargo, la existencia de relaciones no lineales entre los datos pueden limitar la aplicación de estos modelos.[1] En la práctica es muy posible encontrar relaciones no lineales en los datos, tal como sucede en este caso de estudio. Por ello es necesaria la utilización de técnicas capaces de reflejar dicho comportamiento.

La utilización de Redes Neuronales Artificiales (RNAs) para pronósticos de series de tiempo es relativamente nueva en la literatura, sin embargo, lo positivo de los resultados en las aplicaciones prácticas la convierten en una área prometedora.

Para este trabajo, la empresa brindó información histórica de registros mensuales acerca de la utilización de la red de transmisión de los últimos 6 años. Con esta información, se realizó el pronóstico de la demanda para períodos posteriores mediante el uso de RNAs.

Al intentar desarrollar el modelo de RNAs para esta aplicación de pronóstico de series de tiempo, se experimentó y se identificó en la literatura que la exactitud del pronóstico de la RNA depende de varias decisiones críticas en cuanto a la definición de los parámetros que intervienen en el modelo así como de la arquitectura de RNA que se esté utilizando.[2] Algunas de estas decisiones pueden ser tomadas en el proceso de construcción del modelo, mientras que otras requieren ser especificadas antes de que comience la modelación. Sin embargo, no existe una regla establecida que permita tomar varias de estas decisiones de manera adecuada. Por esta razón, en este trabajo se propone y se comprueba mediante el caso práctico una metodología para la selección adecuada de los parámetros de un modelo de RNAs.

Los resultados obtenidos fueron comparados con los que se obtuvieron al analizar las mismas series de tiempo a través de métodos lineales tradicionales, tales como promedios móviles y regresión lineal, entre otros. En los casos analizados el modelo de RNAs construido con la metodología propuesta resultó con mejores resultados, quedando así como una opción viable para la aplicación en la compañía.

ANTECEDENTES

La idea de utilizar RNAs en pronóstico de series de tiempo fue aplicada por primera vez en 1964 cuando Hu utilizó una RNA lineal adaptable de Widrow para el pronóstico del clima.[3] Debido a la ausencia de un algoritmo de entrenamiento para RNA multicapa en el tiempo, la investigación quedó limitada. En 1974 Werbos formuló primero la retropropagación pero no fue conocido por los investigadores en RNAs. A partir de 1986 cuando el algoritmo de retropropagación (del inglés *backpropagation*) fue introducido por Rumelhart *et al*.,[4] el desarrollo de RNAs para pronóstico de series de tiempo ha ido en incremento.[3] Werbos[5] reportó que la RNA entrenada por retropropagación superó el desempeño de los métodos estadísticos tradicionales tales como los procedimientos de regresión y Box-Jenkins en varios casos.

En años recientes, las RNAs han llegado a ser muy populares en el pronóstico de series de tiempo en un gran número de áreas incluyendo finanzas, generación de energía, medicina, recursos del agua y ciencias ambientales, entre otras.[6] Estudios recientes acerca de la aplicación de RNAs en problemas de investigación de operaciones y negocios se pueden encontrar en Zhang[2] y Smith *et al*.[7]

En la mayoría de las aplicaciones realizadas, los autores utilizan RNAs multicapa entrenadas por retropropagación del error para pronósticos a corto plazo y se limitan a utilizar RNAs con una sola neurona en la capa de salida. Sin embargo, en aplicaciones prácticas es común que se desee estimar más de un periodo futuro.

Cuando se desea pronosticar múltiples periodos, muchos investigadores,[3,8-10] han utilizado como recurso un modelo de RNA con una neurona en la capa de salida. Este tipo de RNA se ha utilizado como base para generar pronósticos de múltiples periodos de la siguiente manera: una vez que se tiene el pronóstico para el primer periodo, se itera el modelo

considerando a éste como dato real para calcular el pronóstico del segundo periodo, y así sucesivamente hasta alcanzar el horizonte de planeación deseado. Esta técnica obviamente trae consigo la desventaja de propagar el error de cada uno de los pronósticos a lo largo de todos los periodos que le siguen. Esto es, un mal pronóstico generado en los primeros periodos podría afectar de manera adversa los pronósticos de los últimos periodos.

Una técnica más de RNAs para el pronóstico de múltiples periodos es crear un sólo modelo que simultáneamente genere los pronósticos de múltiples periodos, es decir, una RNA con múltiples salidas (figura 1). Aunque se espera que conduzca a mejores resultados que las técnicas descritas anteriormente,[2] esto aún no ha sido completamente abordado en la literatura. Hay, de hecho, pocas referencias de trabajos desarrollados con la aplicación de esta técnica.

La capacidad de aproximación universal de las RNAs para funciones continuas que tienen primera y segunda derivada en todo su dominio ha sido demostrada matemáticamente. Adicionalmente, varios estudios demuestran que las RNAs pueden aproximar con exactitud diversos tipos de relaciones funcionales complejas.[11-13] Esta última característica es muy importante para la aplicación que aquí se describe, pues de cualquier modelo de predicción se espera que detecte con exactitud la relación funcional entre la variable a predecir y otros factores o variables relevantes.

La combinación de modelación no lineal y aprendizaje a partir de los datos hace que las RNAs

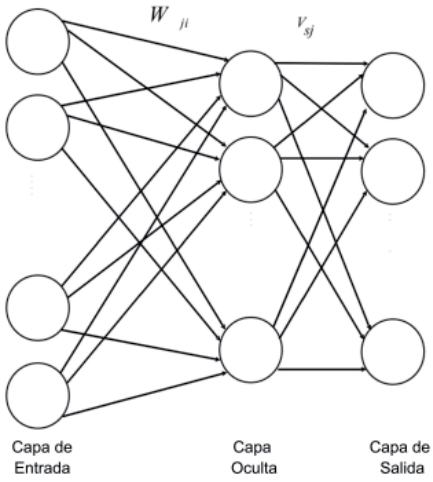


Fig. 1. Red Neuronal Multicapa con múltiples salidas entrenada por retropropagación del error.

sean herramientas flexibles de modelación general atractivas para su aplicación en la realización de pronósticos.

A pesar de que son numerosas las aplicaciones desarrolladas mediante RNAs para pronóstico de series de tiempo y que los resultados han sido satisfactorios, no ha sido posible estandarizar una metodología que garantice la construcción de modelos de RNAs con buen desempeño, entendiendo "desempeño" como la exactitud del pronóstico. Por esta razón, proponemos una metodología con bases de estadística y optimización matemática que permite seleccionar de manera adecuada los parámetros del modelo. Esta metodología se describe a continuación.

METODOLOGÍA PROPUESTA

La figura 2 representa esquemáticamente la metodología que se propone para seleccionar los parámetros del modelo de RNAs para pronóstico de series de tiempo.

La descripción de la metodología es la siguiente:

1) Descripción de la RNA como sistema.
   - Determinar el tipo de RNA que se utilizará para el análisis.
   - Identificar los parámetros controlables.
   - Definir las respuestas de interés (medidas de desempeño del modelo de RNAs).
2) Análisis y diseño de experimentos.
   - Planear, ejecutar e interpretar un diseño estadístico de experimentos.
3) Metamodelación.
   - Describir la superficie de cada respuesta mediante un modelo de regresión apropiado.
4) Problema de optimización.
   - Considerar los metamodelos como funciones objetivo de un problema de optimización.
5) Solución.
   - Resolver los problemas de optimización definidos en el paso anterior. Utilizar múltiples comienzos para escapar de optimalidad local.

Para definir la metodología propuesta, se tomó como base la definición general de "experimento", que es una prueba planeada donde se introducen cambios controlados en un proceso o un sistema con el objetivo de analizar la variación inducida por

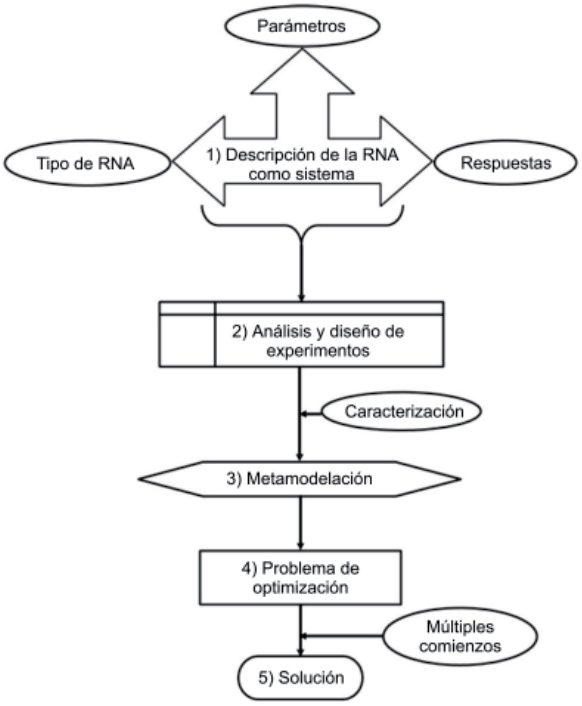


Fig. 2. Metodología para la selección parámetros de un modelo de RNAs.

estos cambios en una medida de desempeño. De esta manera, los factores controlables que intervienen en el experimento corresponden a los parámetros del modelo de RNAs que se desean determinar. Como ejemplos se pueden citar: cantidad de datos de entrada o datos históricos en el caso de series de tiempo; cantidad de neuronas en la capa oculta; algoritmo de entrenamiento; y para el manejo de datos: transformación utilizada y escala de los datos. Así cada corrida experimental indica los valores asignados a los parámetros para construir la RNA correspondiente y bajo los cuales se llevará a cabo el entrenamiento de la misma, y una vez realizado, su validación, para posteriormente cuantificar la calidad de predicción de la RNA a través de las medidas de desempeño seleccionadas para el estudio y su registro como parte del experimento. Generalmente cuando se habla de pronóstico, las medidas de desempeño son medidas de error del pronóstico, por ejemplo el error cuadrado medio o MSE y el error absoluto promedio o MAE.

En los parámetros donde sea posible considerar tres o más valores diferentes, es recomendable utilizar al menos tres de esos valores, con el fin de brindarle curvatura al modelo.

Realizado el experimento se lleva a cabo su análisis con el objetivo de caracterizar la variación producida por los parámetros en las medidas de desempeño del modelo de RNAs. Para ello, requerimos hacer un análisis de varianza basado en un modelo de regresión lineal múltiple de segundo orden con interacciones, similar al de la ecuación (1), bajo el supuesto de que los residuos, ε , son independientes e idénticos y normalmente distribuidos con una varianza desconocida pero constante.

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \beta_{ij} x_i x_j + \varepsilon \qquad (1)$$

La variable dependiente y representa el valor de la medida de desempeño de la RNA, x_i corresponde al valor que toma el parámetro i en cada combinación del experimento, β_0 representa la ordenada al origen del plano de regresión, β_i corresponde al coeficiente de regresión de x_i , β_{ii} es el coeficiente de regresión de x_i^2 y β_{ij} es el coeficiente de regresión de la interacción de entre x_i y x_j ; k es el número de parámetros controlables.

Los coeficientes de regresión típicamente se calculan mediante un procedimiento de reducción de errores cuadrados, disponible en la mayoría de paquetes computacionales comerciales de estadística.

Una vez calculados los coeficientes de regresión, se realiza un análisis de residuos para verificar los supuestos sobre ε así como la adecuación del modelo (1) para representar la medida de desempeño.

Finalmente, se considera el modelo de regresión resultante para cada medida de desempeño como función objetivo de un problema de optimización en el cual se busca encontrar los valores de los parámetros que minimizan el valor de la función objetivo. La formulación P1 muestra la estructura general del problema de optimización.

$$\begin{aligned}
&\textbf{Encontrar} \quad x_i \quad \forall i \in I \quad \textbf{para} \\
&\textbf{Minimizar} \quad z = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \beta_{ij} x_i x_j \\
&\textbf{Sujeto a} \\
&\qquad x_{li} \le x_i \le x_{ui} \qquad \forall i \in I \\
&\qquad x_i \in Z^+ \qquad \forall i \in I \\
&\qquad I = \{1, 2, ..., k\}
\end{aligned} \qquad (P1)$$

En P1, se busca el valor para el i-ésimo parámetro representado por x_i simultáneamente con los valores

de todos los parámetros definidos en el problema para minimizar la función objetivo z, que representa una medida de desempeño.

El problema de optimización resultante es no lineal, la gran mayoría de las veces, y las variables de optimización son enteras, lo cual hace que tal problema sea difícil de resolver. La no convexidad de este problema provoca, además, dificultades para garantizar que la solución encontrada sea una solución óptima global.

Por último en la metodología, para encontrar la solución se resuelven los problemas de optimización de manera independiente con algún optimizador disponible. Si la solución final es la misma para todos los problemas, significa que las medidas de desempeño están correlacionadas pues lo que optimiza a una también optimiza a las otras.

En caso que las soluciones finales encontradas para cada uno de los problemas de optimización sean diferentes, se deberá utilizar una técnica de optimización multicriterio para ofrecer un abanico de soluciones que representen los mejores compromisos entre las medidas de desempeño de la RNA. A estas soluciones se les llama "eficientes". De este conjunto de soluciones, el tomador de decisiones elige una que convenga a sus intereses.

La solución a la que se llega a través de esta metodología establece los parámetros del modelo de RNAs que habrán de utilizarse para el pronóstico, de tal manera que el desempeño de predicción sea competitivo.

Para probar la metodología propuesta se realizó el estudio de un caso práctico en una compañía local, el cual se presenta a continuación.

CASO DE ESTUDIO Y RESULTADOS

Se analizaron dos series de tiempo, representadas en las figuras 3 y 4, con un horizonte de planeación de doce meses y utilizando un modelo de RNA como el de la figura 1. Se preprocesó la información transformando los datos en la escala de [-1, 1]. El algoritmo de entrenamiento utilizado para la RNA fue el de Levenberg-Marquart con múltiples comienzos.

En este caso se consideraron como parámetros controlables la cantidad de datos históricos que se utilizarían para generar el pronóstico (lags) y la cantidad de neuronas y en la capa oculta. Entonces, para efectos de la metodología, x_1 representa al parámetro lags y x_2 corresponde al parámetro neuronas.

Para crear los conjuntos de entrenamiento y validación se utilizó una distribución uniforme que ayudó a seleccionar aproximadamente el 70% de los patrones disponibles para el entrenamiento y el resto para la validación. El desempeño de la RNA fue cuantificado mediante el Error Cuadrado Medio o MSE, considerando a éste tanto para la fase de entrenamiento (MSE_T) como para la fase de validación (MSE_V) del modelo.

Para el entrenamiento del modelo de RNAs consideramos múltiples inicios ya que al hacer la actualización de pesos mediante la retropropagación del error realmente se está minimizando una función de error, que es no lineal y no convexa. Por esta razón, el punto de inicio de la optimización es determinante para los pesos finales que adquieren las conexiones de la RNA una vez que el entrenamiento termina.

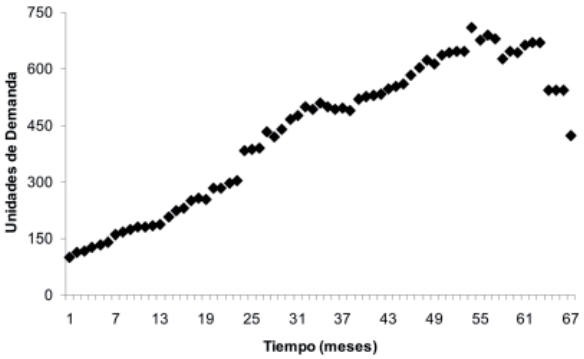


Fig. 3. Comportamiento de la demanda, Serie 1.

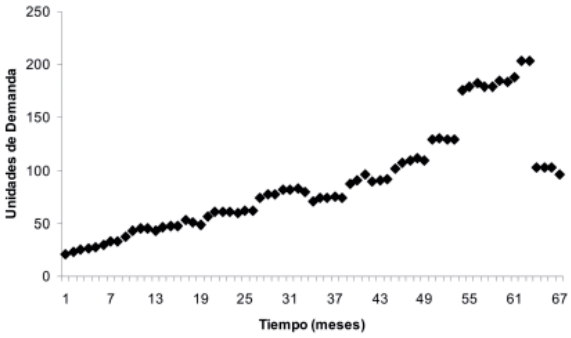


Fig. 4. Comportamiento de la demanda, Serie 2.

Los pesos se fijaron inicialmente todos en -1, -0.5, 0, 0.5 y 1, así como una vez más de manera aleatoria. Por tanto, para cada corrida experimental se crearon seis RNAs diferentes. De éstas se seleccionó el modelo con menor error de validación, para su registro en la tabla experimental.

En la experimentación con la Serie 1 se varió lags en el rango [3, 15] y neuronas en el rango [2, 10]. Los valores específicos que se consideraron para cada factor en su nivel correspondiente fueron: lags = {3, 6, 9, 12, 15} y neuronas = {2, 4, 6, 8, 10}. En el experimento se utilizó un diseño factorial que resultó en un total de 25 combinaciones para correr el modelo de RNAs. Cuando se obtuvieron los metamodelos se observó que el porcentaje de variación explicado por los metamodelos era muy bajo, lo que significa que no eran buenas aproximaciones. Con esta información, se decidió enfocar (reducir) el área experimental con el fin de encontrar metamodelos confiables.

Gracias a un análisis gráfico se determinó que se podían considerar potencialmente competitivos los modelos de RNAs con los parámetros lags y neuronas dentro de sus tres primeros niveles. Se tomaron entonces las corridas experimentales resultantes de la combinación de lags = {3, 6, 9} y neuronas = {2, 4, 6}. Se realizó nuevamente el análisis de varianza y en esta ocasión los metamodelos resultaron apropiados así que se pasó a la optimización.

Al resolver los problemas resultantes y utilizando múltiples soluciones iniciales, ambos metamodelos llevaron a soluciones óptimas que gráficamente se pudieron corroborar como globales respectivamente. Sin embargo, la solución óptima para el MSE_T fue distinta a la que se obtuvo al minimizar el MSE_V. Para el primero, la optimización llevó a un modelo de RNA con 7 datos históricos (lags) y 5 neuronas en la capa oculta; para el segundo, el modelo de RNA con mejor desempeño fue aquél en el que se consideraban 5 datos históricos y 5 neuronas en la capa oculta. Estos resultados indicaron que los objetivos estaban en conflicto.

Dada la importancia de obtener modelos de RNA con buena calidad de aproximación y generalización, se decidió darle mayor importancia al MSE_V. Así que la solución final fue (5,5), es decir, 5 datos históricos y 5 neuronas ocultas.

Ya que se tuvo la solución final, es decir, los valores a los cuales debían ser ajustados los parámetros, se construyó el modelo de RNA correspondiente para la realización del pronóstico. Se consideraron nuevamente las seis diferentes inicializaciones de pesos en las conexiones y se seleccionó la RNA con menor valor de MSE_V. En la figura 5 se muestra el pronóstico que se obtuvo con este modelo de RNA, así como los datos reales y el pronóstico obtenido por el método de regresión lineal. Se presenta aquí este último método por reportar el mejor desempeño de pronóstico basado en MSE con esta serie de entre ocho técnicas tradicionales: promedio móvil para n=5, 8 y 10, Arima(0,1,1), suavizado exponencial simple, regresión lineal, Arima(0,2,2) y suavizado exponencial doble.

Un proceso similar se realizó en el análisis de la Serie 2, el modelo de RNA con mejor desempeño fue aquél con 5 datos históricos y 3 neuronas en la

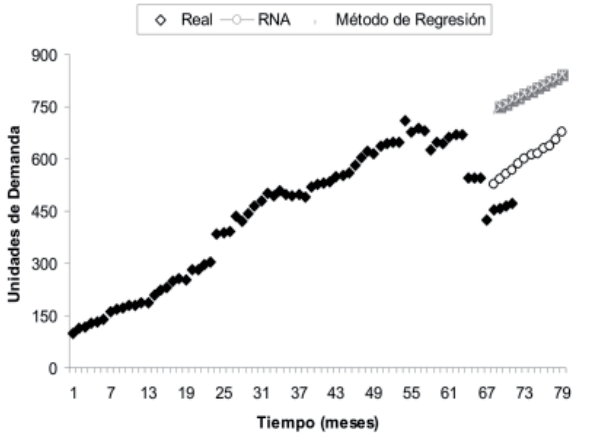


Fig. 5. Demanda real, pronósticos por RNA y por el método de regresión para la serie 1.

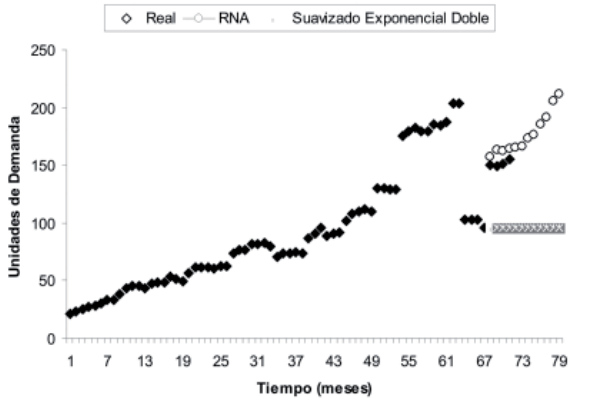


Fig. 6. Demanda real, pronósticos por RNA y por el método de suavizado exponencial doble para la serie 2.

capa oculta. La figura 6 muestra gráficamente la comparación entre los datos reales, el pronóstico obtenido por suavizado exponencial doble, así como por el modelo de RNAs. El suavizado exponencial doble fue el más competitivo de la lista de técnicas tradicionales detallada anteriormente.

Como se puede apreciar en ambos casos, las RNAs construidas con el método propuesto reportaron mejor desempeño de pronóstico.

CONCLUSIONES

En este trabajo se propuso una metodología de selección de parámetros de un modelo de RNAs que utiliza técnicas establecidas y confiables y hace entendible la interrelación entre los varios parámetros de la RNA. Se demostró el funcionamiento de la metodología a través de un caso práctico, en el que se utilizaron modelos de RNA con múltiples salidas.

Los resultados de este trabajo apoyan la utilización de las RNAs como técnicas confiables de pronóstico y apuntan a la factibilidad de su instauración en la industria.

Como extensiones de este trabajo se plantea comparar el método propuesto con más técnicas tradicionales de pronóstico, así como otros métodos de construcción de RNAs.

AGRADECIMIENTOS

Los autores agradecen al CONACYT, la FIME y la UANL por las becas recibidas para los estudiantes involucrados en este trabajo. Agradecen también las aportaciones de Ma. Guadalupe Villarreal Marroquín, apoyada por el proyecto UANL-PAICYT CA 1069-05.

REFERENCIAS

1. Makridakis S., Anderson A., Carbone R., Fildes R., Hibbon M., Lewandowski R., Newton J., Parsen E., and Winkley R., "The accuracy of extrapolation (time series) methods: Results of a forecasting competition", Journal of Forecasting, 1982,Vol. 1, pag. 111-153.
2. Zhang G. P., Neural Networks in Business Forecasting, Idea Group Publishing, Georgia State University, EUA, 2004.
3. Zhang G., Patuwo E., and Hu Y. M., "Forecasting with artificial neural networks the state of the art", International Journal of Forecasting, 1998, Vol.14, No. 1, pag. 35-62.
4. Rumelhart D-E., Hinton G. E., and Willians R. J., "Learning representartions by backpropagating errors", Nature, 1986, 323 (6188), pag. 533-536.
5. Werbos P. J., "Generalization of backpropagation with applications to a recurrent gas market model", Neural Networks, 1988, Vol. 1, pag. 339-356.
6. Maier H. R., and Dandy G. C., "Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications", Environment Modelling & Software, 2000, Vol. 15, pag. 101-124.
7. Smith K. A. and Gupta JND, "Neural networks in business: techniques and applications for the operations researcher", Computers and Operations Research, 2000, Vol. 27, Num. 11-12, pag.1023-1044.
8. Hwarng H. B., "Insights into neural-network forecasting of time series corresponding to ARMA (p,q) structures", Omega: The International Journal of Management Science, 2001, Vol. 29, No. 3, pag. 273-289.
9. Hill, T., W. Remus, and M. OπConnor, "Neural Network Models for Time Series Forecasts", Management Science, 1996, Vol. 42, Num. 7, pag. 1082-1092.
10. Nelson, M., T. Hill, W. Remus and M. O'Connor, "Time Series Forecasting Using Neural Networks: Should the Data Be Deseasonalized First?", Journal of Forecasting, 1999, Vol.18, Num.5, pag. 359-370.
11. Irie B., and Miyake S., "Capabilities of three-layered perceptrons", Proccedings of the IEEE International Conference on Neural Networks I, 1988, pp. 641-648.
12. Hornik K., Stinchcombe M., and White H.,"Multilayer feedforward networks are universal approximators", Neural Networks, 1989, Vol. 2, No. 5, pag. 359-366.
13. Cybenko G., "Approximation by superpositions of sigmoidal function", Mathematical Control Signals Systems, 1989, Num. 2, pag. 303-314.