Las redes neuronales artificiales como herramienta para la minería de datos en un contexto financiero

Pedro J. García-Laencina, José L. Roca-González, M. Ángeles Varela-Jul,
Carmen de Nieves-Nieto, Joaquín Roca-Dorda
Centro Universitario de la Defensa (CUD) de San Javier, MDE-UPCT
Base Aérea de San Javier, C/ Coronel López Peña, s/n 30720 Santiago de la Ribera (Murcia) España
E-mail: pedroj.garcia@cud.upct.es

Resumen. *Las técnicas de aprendizaje máquina, entre ellas destacan las redes neuronales artificiales, han sido utilizadas con éxito en multitud de contextos y problemas de carácter financiero y económico. Este trabajo resume el procedimiento general para la minería de datos (extracción de conocimiento a partir de un conjunto de datos que define el problema bajo estudio) y resuelve un caso real de predicción de variables financieras utilizando novedosos algoritmos para el diseño de redes neuronales artificiales. Los resultados muestran las ventajas de estos nuevos modelos computacionales en términos de precisión en la predicción y tiempo de cálculo.*

1 Introducción

Actualmente, las empresas, organizaciones e instituciones manejan cantidades enormes de información que se encuentran almacenadas en bases de datos de gran tamaño. Sin embargo, la información "en bruto" (*raw information*) de las bases de datos carece de utilidad para las empresas y, por ello, es muy necesario extraer conocimiento, es decir, convertir la información en su estado primario a información útil [1]. Por ejemplo, las empresas innovadoras están empleando sus bases de datos para localizar clientes de gran valor (*high-value customers localization*), cambiar las ofertas de sus productos con objeto de incrementar su volumen de ventas (*product assessment*), reducir pérdidas derivadas de posibles impagos (*non-payment prediction*), etc.

Las técnicas empleadas para la extracción de conocimiento de las bases de datos están integradas dentro de dos áreas interrelacionadas entre sí: la minería de datos y el aprendizaje máquina. La *Minería de Datos* (*Data Mining*, DM) se define como la extracción no trivial de información implícita, previamente desconocida y potencialmente útil, a partir de datos [1]. Por otro lado, el *Aprendizaje Máquina* (*Machine Learning*, ML), también denominado aprendizaje automático o computacional, es una rama de la Inteligencia Artificial que trata de construir sistemas computacionales que optimicen un criterio de rendimiento utilizando datos o experiencia previa asociado a un determinado problema [2]. Durante el proceso de optimización, se dice que la maquina aprende o modela el problema objetivo. El sistema computacional obtenido (*máquina* o *modelo*) transforma los datos en conocimiento y proporciona modelos de propósito general que se adaptan a las circunstancias. Así, de manera general, se puede decir que las técnicas empleadas en DM y ML consisten en, dado un conjunto de datos (numéricos y/o nominales) referidos a un problema cualquiera, aplicar técnicas y algoritmos de inteligencia artificial que realizan un aprendizaje automático para establecer patrones y modelos sobre esos datos y así extraer conclusiones sobre ellos. La Fig. 1 muestra el esquema general utilizado para la extracción de conocimiento utilizando DM y ML. Se pueden distinguir dos etapas principales: *Diseño* y *Operación*. Primero, en la etapa de Diseño, el problema objetivo viene definido por un conjunto de datos (matriz de N filas y n columnas), que es dividido de dos subconjuntos: entrenamiento y test. El conjunto de entrenamiento se emplea para el diseño de la máquina o modelo (es decir, se extrae el conocimiento del conjunto de entrenamiento), mientras que los datos de test se utilizan para evaluar las prestaciones de la máquina obtenida durante el entrenamiento. Una vez que se dispone de un buen modelo que resuelve el problema objetivo, este máquina puede ser empleada en la etapa de Operación para obtener/predecir la información deseada a partir de un conjunto de datos nuevos que no han sido utilizados durante la etapa de Diseño.

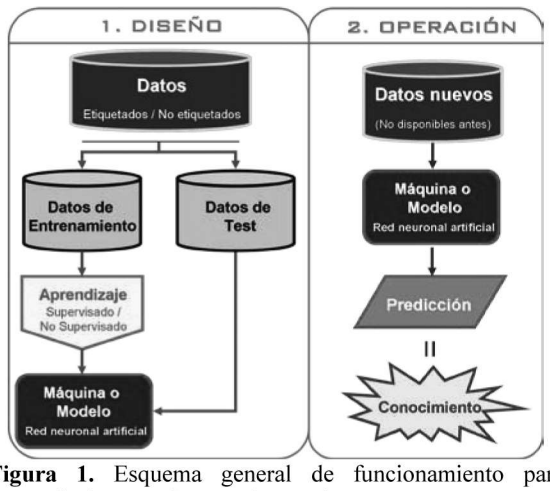

Figura 1. Esquema general de funcionamiento para Aprendizaje Máquina y Minería de Datos, compuesto por dos etapas principales de Diseño y Operación.

Un ejemplo es el sistema empleado por las entidades financieras para la concesión o denegación automática de créditos, conocido como '*credit*

scoring'. Este sistema asigna el riesgo de la operación planteada -cantidad y tipo de crédito- a partir de una serie de datos (n variables) del solicitante, como son ingresos, profesión, edad, ahorros, historial de operaciones, etc. El conjunto de entrenamiento está constituido por la base de datos de N anteriores solicitudes de crédito que contiene la información de los N clientes, mientras que la información objetivo se encuentra en una variable binaria que indica si un determinado cliente no acabó saldando la deuda solicitada. A partir de estos datos (conjunto de diseño) es posible construir un modelo que determine la probabilidad de que un nuevo cliente no haga frente al préstamo solicitado.

2 Un caso de estudio: Modelización de la rentabilidad de los activos

Con objeto de ilustrar las distintas etapas necesarias para la minería de datos, este trabajo utiliza un problema real cuyo objetivo es modelar la rentabilidad de los activos (ROA, *Return on Assets*) a partir de 29 variables financieras de 200 empresas industriales desde 1991 hasta 1995. La siguiente tabla muestra la lista completa de los 29 indicadores utilizados para modelar el ROA.

Tabla 1. Variables financieras para modelar el ROA

Variable	Significado
x_1	Sector de la industria
x_2	Número de acciones vendidas en el año
x_3	Número de empleados
x_4	Ratio de volumen de ventas de seguridad
x_5	dividendo por acción por año
x_6	Ventas
x_7	Otros activos
x_8	Provisión sobre otros activos
x_9	Propiedades y equipos
x_{10}	Provisiones sobre propiedades y equipos
x_{11}	Activos fijos
x_{12}	Stocks o inventarios
x_{13}	Deudas por cobrar
x_{14}	Banco y Caja
x_{15}	Total de activos disponibles
x_{16}	Total del capital de la empresa
x_{17}	Deudas
x_{18}	Deuda financiera
x_{19}	Deudas a corto plazo
x_{20}	Deudas a largo plazo
x_{21}	Deudas totales
x_{22}	Costes de los trabajadores
x_{23}	Dotaciones en amortizaciones
x_{24}	Beneficio antes de impuestos
x_{25}	Impuestos sobre los intereses
x_{26}	Ingresos financieros
x_{27}	Beneficio antes de impuestos + Ingresos financieros
x_{28}	Ingresos extraordinarios
x_{29}	Impuesto sobre sociedades

El ROA es la tasa que mide la rentabilidad de un proyecto o de una sociedad respecto de los activos totales y es un factor que los mercados consideran fundamental para la valoración de las empresas.

Inicialmente, la base de datos (conjunto de diseño) consta de un total de 650 casos. Este conjunto es dividido aleatoriamente en dos partes: un conjunto de entrenamiento con un 70% de los casos y un conjunto de test formado por el 30% restante. Esta manera de proceder está muy extendida ya que es necesario disponer de un número suficiente de casos para que el conjunto de entrenamiento sea representativo.

2.1 Modelo predictivo mediante redes neuronales artificiales

Las Redes Neuronales Artificiales (ANNs, *Artificial Neural Networks*) fueron originalmente una simulación abstracta de los sistemas nerviosos biológicos, formados por un conjunto de unidades llamadas "*neuronas*" conectadas unas con otras. Estas conexiones o pesos tienen una gran semejanza con las dendritas y los axones en los sistemas nerviosos biológicos. La investigación en ANNs ha experimentado una extensa actividad desde sus orígenes en el año 1941 hasta la fecha actual, aunque fue a partir de la década de los 80 cuando se incremento considerablemente el interés en estos modelos computacionales al demostrarse la capacidad de aproximar cualquier problema con un número suficiente de casos y neuronas. El Perceptron Multicapa (MLP, *Multi-Layer Perceptron*) es el modelo más típico de las ANNs. Un MLP está definido por un conjunto de vectores de pesos que conectan una capa de entrada con una capa de salida utilizando una capa de H neuronas ocultas entre ambas. El algoritmo de diseño más utilizado se denomina retro-progragación (BP, Back-Propagation) y está basado en un proceso de optimización basado en el cálculo de los gradientes de la función de error dada por las salidas obtenidas por el MLP –obtenidos al procesar los casos de entrenamiento- y las salidas deseadas. La Fig. 2 muestra la arquitectura básica de un MLP con n variables de entrada, H neuronas ocultas y una unidad de salida (variable objetivo).

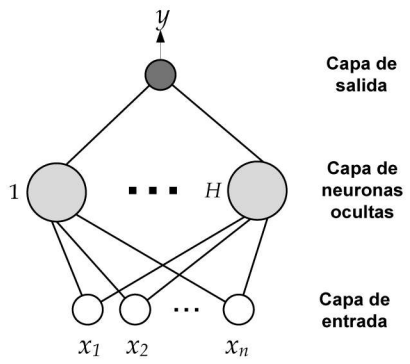

Figura 2. Arquitectura básica de un MLP

Sin embargo, presentan diversos inconvenientes durante el proceso de optimización de los hiper-parámetros (i.e., pesos y número de neuronas) como

son el elevado tiempo computacional requerido y la convergencia a soluciones sub-optimas (mínimos locales) para un determinado problema al usar técnicas de optimización basados en gradiente [1]. Recientemente, el algoritmo de entrenamiento conocido como *Extreme Learning Machine* (ELM) [2] ha permitido solventar estos claros inconvenientes que presentan las técnicas tradicionales de entrenamiento de ANNs.

A continuación, se procede a modelar la variable ROA utilizando un MLP entrenado con el tradicional algoritmo BP (con gradiente conjugado escalado) y el novedoso ELM. En particular, se ha realizado un barrido desde 1 a 75 neuronas ocultas y se han hecho un total de 30 simulaciones para obtener resultados promedios. La Fig. 3 muestra los resultados obtenidos en términos del RMSE (*Root Mean Square Error*) para el conjunto de test. Como se puede apreciar en esta figura, el algoritmo ELM consigue un mejor modelo con un número de neuronas superior a 15. En concreto, la mejor predicción, **0.791±0.021** (media y desviación estándar), se consigue con una red ELM de 41 neuronas. Por otro lado, la Fig. 4 muestra el tiempo de entrenamiento utilizado en ambos algoritmos de aprendizaje. En este caso, podemos comprobar que el algoritmo ELM es 10 veces más rápido que el tradicional algoritmo BP.

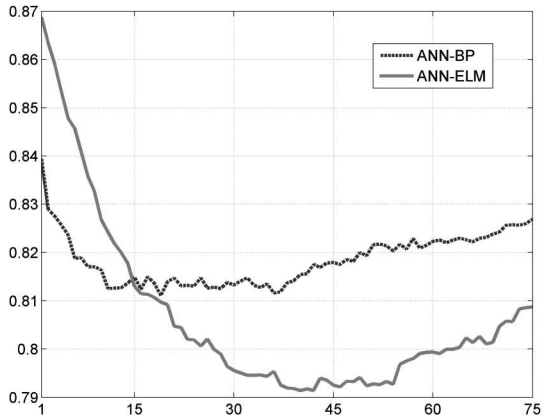

Figura 3. Evolución del RMSE en el conjunto de test con respecto del número de neuronas.

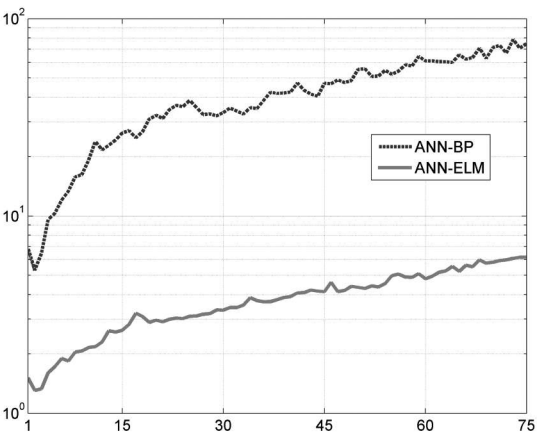

Figura 4. Evolución del tiempo de cómputo (entrenamiento) con respecto del número de neuronas.

Hay que destacar que el problema ha sido modelado utilizando los 29 indicadores financieros, sin embargo es muy probable que no todas las variables resulten útiles/relevantes para predecir el ROA. Por ello, a continuación, se realiza una búsqueda exhaustiva incremental de aquellas variables que resultan relevantes para aprender el ROA mediante una arquitectura neuronal basada en el algoritmo ELM. En concreto, esta búsqueda incremental actúa de forma iterativa, añadiendo en cada paso la variable que, incorporada a las que ya han sido escogidas, proporciona el error mínimo de validación cruzada de tipo *leave-one-out*. Las características seleccionadas son las siguientes: x_{27}, x_{29}, x_{21}, x_{13} y x_{15}. Estas variables han sido ordenadas incrementalmente según su importancia para predecir el ROA, por lo que el indicador financiero más relevante es x_{27} "Beneficio antes de impuestos + Ingresos financieros". El modelo obtenido proporciona un error de **0.765±0.016** (media y desviación estándar de 30 simulaciones) y el número medio de neuronas es igual a 23, que es menor que el tamaño de la mejor arquitectura ELM con todas las variables de entrada. Por tanto, la correcta selección de variables relevantes para modelar el ROA consigue una mejor predicción con una arquitectura neuronal más sencilla que en el problema original.

3 Conclusiones

Este trabajo ha introducido los conceptos fundamentales de la minería de datos y las redes neuronales artificiales para su aplicación en contextos financieros y económicos. En particular, se ha estudiado un problema real cuyo objetivo es modelar la rentabilidad de un conjunto de empresas a partir de una base de datos formada por 29 indicadores financieros. Los resultados obtenidos muestran las ventajas de emplear las últimas técnicas desarrolladas para el entrenamiento y diseño de redes neuronales artificiales. La información proporcionada (la predicción del ROA y el subconjunto de indicadores financieros más importantes) por la red neuronal puede servir como sistema de ayuda a la decisión para un analista financiero.

Referencias

[1] P Giudici. *Applied data mining: Statistical methods for business and industry*. Wiley, West Sussex, England. (2003).

[2] S. Haykin. *Neural networks: A comprehensive foundation*. Prentice Hall, New York, USA. (1998).

[3] P. J. Garcia-Laencina, A. Bueno-Crespo, J.-L. Sancho-Gómez. *Design and Training of Neural Architectures using Extreme Learning Machines*. Neurocomputing: Learning, Architectures and Modeling. Nova Science Publishers. 2011. In press.