

PREDICCIÓN DEL ÉXITO EN ESTUDIOS UNIVERSITARIOS MEDIANTE REDES NEURONALES

Lourdes Molera
M^a Victoria Caballero
Universidad de Murcia

ABSTRACT

Con este trabajo se pretende estudiar la posibilidad que tiene un alumno matriculado en la Universidad de Murcia (en la licenciatura de Economía o ADE, o en la diplomatura de Ciencias Empresariales) de finalizar con éxito sus estudios en un determinado periodo de tiempo. Para ello, se construye una red neuronal a partir de información acerca de sus características socioeconómicas y sus resultados académicos.

Las redes neuronales constituyen una nueva técnica no paramétrica de análisis de datos multivariante, que ha sido utilizada recientemente en el ámbito de la Economía de la Educación. Esta herramienta es más flexible y permite formular relaciones más complejas que las técnicas estadísticas tradicionales.

Palabras clave: redes neuronales, educación universitaria

1. INTRODUCCIÓN

Las redes neuronales son modelos matemáticos que intentan imitar la estructura y funcionamiento del cerebro. Una red neuronal está formada por un conjunto de elementos simples interconectados, que es capaz de procesar la información disponible para aprender y así clasificar, predecir, discriminar...

Una diferencia esencial entre las redes neuronales y otros modelos matemáticos, que se usan con más frecuencia, reside en el proceso de construcción, puesto que en estos últimos es necesario especificar *a priori* la función que sigue cualquier proceso que se pretenda modelizar, mientras que una red neuronal desarrolla una aproximación a la relación funcional desconocida que liga unas variables con otras.

En este trabajo se distinguen dos secciones principales. En la primera se realiza una introducción sintética a las redes neuronales con conexiones hacia delante que utilizan el aprendizaje de retropropagación supervisado haciendo uso del algoritmo del gradiente descendente. En la segunda sección se aplica este tipo de redes a unos ficheros de datos que recogen información sobre alumnos que comenzaron sus estudios universitarios (licenciaturas en economía y ADE o diplomatura en empresariales) en el curso 94-95.

2. REDES NEURONALES

Una red neuronal está formada por unidades simples llamadas nodos o neuronas que se disponen en capas y se relacionan entre sí por conexiones. Se distinguen tres tipos de capas, una capa de entrada, también llamada capa *input*, otra capa de salida o capa *output*, y una o varias capas ocultas. Hay distintos tipos de redes neuronales, pero en este trabajo vamos a tratar únicamente con aquellas que procesan la información hacia delante, es decir, las neuronas de cada capa sólo mantienen conexiones con todas las neuronas de la capa siguiente. A este tipo de redes se las llama *feed-forward*.

La información que recibe un nodo de una capa oculta o de salida es el resultado de ponderar la salida de cada nodo de la capa previa y agregarla (generalmente el proceso de agregación consiste en la suma lineal). A dicha información agregada se le aplica una función de activación para obtener la salida de ese nodo, que pasará a ser información de entrada para los nodos de la capa posterior o será el resultado final si se trata de la capa de salida. Por ejemplo, consideremos el nodo j de la capa $k+1$. Este nodo está conectado con los nodos de la capa previa k , donde se puede suponer que hay 3 nodos, cuya información de salida es x_{k1} , x_{k2} y

x_{k3} , que se pondera por unos pesos ω_{1j} , ω_{2j} , ω_{3j} , y se agrega, obteniéndose $\sum_{i=1}^3 \omega_{ij} x_{ki}$. A esta

información se le aplica una función de transferencia o activación f , y se tiene

$o_{k+1,j} = f\left(\sum_{i=1}^3 \omega_{ij} x_{ki}\right)$ como información de salida del nodo j de la capa $k+1$. Las funciones de

activación tienen el propósito de controlar la señal de salida de cada nodo, excepto en la capa de entrada, siendo las funciones de activación más comunes la sigmoide, la tangente hiperbólica, la gaussiana o la secante hiperbólica.

Atendiendo al proceso de aprendizaje, se va a utilizar un tipo de red de aprendizaje supervisado, puesto que dentro del conjunto de datos que constituyen la información para la red, hay un subconjunto que sirve para comparar el resultado obtenido por ésta. Cuando los resultados obtenidos por la red no pueden ser comparados con un subconjunto de datos, que serían los deseados, se dice que el aprendizaje es no-supervisado.

2.1. Redes neuronales feed-forward back-propagation

Describiremos este tipo de red, el algoritmo y la técnica que se utiliza con un ejemplo. Supongamos que nuestra red va a estar formada por tres capas, una capa input, otra capa output y sólo una capa oculta. La capa input tiene cinco nodos, cada uno de los cuales incorpora una componente del vector de entrada. Los nodos de la capa oculta son tres y la capa output sólo tiene un nodo (sólo un posible resultado, pero podría tener más). Se llama w_{ij} a la ponderación entre el nodo input i y el nodo oculto j y v_j a la ponderación entre el nodo oculto j y el nodo de la capa output. Se denota por H_j el resultado obtenido por el nodo oculto j y por O el output final. El input del nodo oculto j es el valor agregado $\sum_{i=1}^5 w_{ij} x_i$ y su output, utilizando como función de activación la sigmoide, $H_j = 1/(1 + \exp(-\sum_{i=1}^5 w_{ij} x_i))$. De ahí a la capa output se produce otro proceso de agregación, y nuevamente se aplica la función de activación elegida, obteniéndose $O = 1/(1 + \exp(-\sum_{j=1}^3 v_j H_j))$. Este output final se compara con el resultado deseado, que llamaremos Y .

Se ha descrito una red neuronal muy sencilla; las redes se pueden complicar mucho más incrementando el número de capas ocultas y el de nodos en cada una de ellas, así como el número de nodos en la capa output.

2.1.1. Número de capas ocultas y su número de nodos

En la construcción de una red neural hay que resolver inicialmente dos cuestiones: ¿cuál es el número adecuado de capas ocultas? y ¿cuántos nodos debe de haber en la/s capa/s

oculta/s?. Respecto de la primera pregunta, se puede contestar haciendo uso del teorema de Kolmogorov que dice que cualquier función continua creciente en n variables puede ser computada usando sólo sumas lineales y una función no lineal continua y creciente en una variable, lo que demuestra que el conjunto de redes neuronales con tres capas es denso en el espacio de todas las funciones continuas en n variables. Como consecuencia, una red neuronal con tres capas puede aproximar cualquier función (Kurkova, 1992).

Respecto del número de nodos en la capa oculta, la teoría existente no nos dice cuál debe ser. La elección de cuántos nodos depende de distintos factores, como la naturaleza del problema o el tamaño y calidad de los datos (Brockett y otros, 1997). Dos o menos nodos ocultos puede simplificar demasiado la red y no tener suficiente capacidad para aprender. Por el contrario, demasiados nodos en la capa oculta puede dar lugar a sobreaprendizaje, es decir, buen ajuste a los datos de aprendizaje pero escaso poder predictivo para nuevos datos (Tam y Kiang, 1992).

Decidir el número de capas y de nodos, así como las variables inputs y outputs, determina lo que se denomina arquitectura o topología de la red neuronal.

2.1.2. Algoritmo back-propagation

Una vez decidida la arquitectura de la red, se asignan unos valores iniciales a las ponderaciones de modo aleatorio. Se presentan a la red los valores de las variables inputs correspondientes a un individuo y estos datos se van propagando a través de todas las capas hasta obtener un resultado output, que se compara con el resultado deseado. Este proceso se lleva a cabo con los datos correspondientes a todos los individuos disponibles, lo que constituye una época. Se calcula el error cometido en cada época como la suma de los cuadrados de las diferencias entre output y resultado deseado para cada elemento de la muestra. El proceso de aprendizaje consiste en ir modificando los pesos época tras época de modo que el error sea mínimo.

Básicamente el algoritmo de back-propagation (Rumelhart y otros, 1986) consiste en transmitir el error hacia atrás en la red partiendo de la capa de salida. Este proceso se transmite al resto de las capas, permitiendo actualizar los pesos de acuerdo con las contribuciones de cada nodo. Con esta actualización se vuelve a aplicar la red a los datos disponibles y se vuelve a medir el error, poniéndose en marcha nuevamente el mecanismo hacia atrás. El proceso continuará hasta que el error sea menor que un cierto umbral previamente fijado o bien porque se haya llegado al máximo de iteraciones.

La técnica que se utiliza en la aplicación de este algoritmo es la del gradiente descendente. Retomando el ejemplo que se ha puesto para describir la arquitectura de una red, realizaremos someramente una descripción matemática de esta técnica. El error de toda la red

viene dado por $E = \frac{1}{2} \sum_{t=1}^T (Y_t - O_t)^2$, donde T es el número de elementos de la muestra.

Recordamos que la red sólo tenía un nodo en la capa de salida y se llamaba O al resultado obtenido e Y al resultado deseado.

1. Primeramente se actualizan los pesos de las conexiones de la capa oculta a la capa output. Se necesita para ello el vector gradiente del error respecto de los v_j ,

$$\frac{\partial E}{\partial v_j} = - \sum_t (Y_t - O_t) \frac{\partial O_t}{\partial v_j} = - \sum_t (Y_t - O_t) O_t (1 - O_t) H_{jt}.$$

El valor de los pesos v_j se actualiza con una tasa negativa, por lo que el nuevo valor vendría dado por

$$v_j^* = v_j + (-\eta) \frac{\partial E}{\partial v_j}.$$

2. En segundo lugar, se actualizan los pesos de las conexiones de la capa input a la capa oculta, teniendo en cuenta la expresión

$$\frac{\partial E}{\partial w_{ij}} = - \sum_t (Y_t - O_t) O_t (1 - O_t) v_j H_{jt} (1 - H_{jt}) x_{it}.$$

Para que el proceso de aprendizaje se realice más rápidamente, y a la vez reducir las posibilidades de convergencia a un mínimo local, se puede añadir un término de momento, que recuerda el cambio realizado en el paso anterior (Curram y Mingers, 1994).

El hecho de repetir este proceso durante muchas épocas puede dar a lugar a sobreaprendizaje. Para evitarlo, se suele dividir el conjunto de datos muestrales en dos subconjuntos, uno destinado a entrenamiento y otro a validación (Curram y Mingers, 1994). Con los datos de entrenamiento (al menos el 60% del total) se actualizan los pesos, siguiendo el algoritmo backpropagation ya descrito. La red obtenida con los pesos actualizados cada época se pasa sobre los datos del conjunto de validación y medimos el error cometido. Si este error aumenta a lo largo de las épocas debemos parar el entrenamiento para evitar el sobreaprendizaje.

En general, se suele considerar también un tercer subconjunto de datos de test, destinado a contrastar el poder de generalización de la red asociada a las ponderaciones finales.

3. APLICACIÓN EMPÍRICA

3.1. Descripción de los datos

Se dispone de datos de 447 alumnos matriculados por primera vez en las licenciaturas de Economía (152) y ADE (295) en el curso 1994-95, y también de datos de 391 alumnos pertenecientes a la diplomatura de Empresariales. Esta información hace referencia a sus

características socioeconómicas (sexo, edad, estudios del padre, estudios de la madre, trabajo del padre, trabajo de la madre), así como académicas (nota de acceso, tipo de acceso, tipo de centro de procedencia –privado o público- y fecha de finalización de estudios universitarios).

Con esta información se pretende aplicar la técnica de las redes neuronales tanto a los alumnos matriculados en una de las licenciaturas como a los de la diplomatura, y así, dadas las características socioeconómicas y académicas preuniversitarias de un alumno poder conocer *a priori* si tendrá éxito o no en estos estudios universitarios, dando una medida del error cometido en esta clasificación. Entendemos por éxito el hecho de que un alumno matriculado en una de las dos licenciaturas finalice sus estudios en los cursos 97-98, 98-99 ó 99-00, mientras que para un alumno de la diplomatura en los cursos 96-97, 97-98 ó 98-99. El motivo de escoger esta manera de medir el éxito se debe al escaso porcentaje de alumnos que acaban en un período de tiempo menor.

Las variables input para la red correspondiente a los alumnos de la diplomatura son las siguientes:

- EDAD: variable numérica que mide la edad del alumno en 1994.
- ESTMAD: variable categórica con tres niveles para los estudios de la madre (sin estudios o primarios, secundarios y universitarios).
- NOTAACC: nota de acceso a los estudios universitarios considerados.
- CUPO: variable categórica con 5 niveles que representa la procedencia del alumno (titulado, FP, COU A o bachillerato científico tecnológico, COU B o bachillerato biosanitario, COU C o bachillerato de ciencias sociales).

En el caso de la red correspondiente a los datos de ambas licenciaturas, las variables de entrada son:

- TITULAC: variable dicotómica que distingue entre las licenciaturas de Economía y ADE.
- TRAMAD: variable dicotómica que toma el valor 1 si la madre trabaja fuera de casa y 0 en caso contrario.
- NOTAACC: nota de acceso a los estudios universitarios considerados.
- VIA: variable dicotómica que vale 1 para los alumnos procedentes de COU A o bachillerato científico-tecnológico y 0 en caso contrario.
- CENTRO: variable dicotómica que toma el valor 1 si el centro del que procede el alumno es privado y 0 si es público.

En ambos casos, la elección de estas variables se debe a que eran las más significativas según el criterio de la χ^2 o en la correspondiente regresión logística.

3.2. Construcción de las redes neuronales y resultados obtenidos

Las redes que utilizamos sólo tienen una capa oculta y el número de nodos en ella se ha elegido de manera que se obtenga la mejor clasificación tanto sobre el conjunto de entrenamiento como el de test. Destinamos el 70% de los datos al entrenamiento de la red, el 10% a la validación y los restantes para testar.

La función de activación elegida es la tangente hiperbólica,

$$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

puesto que el proceso de convergencia se realiza más rápido que con la sigmoide, y la técnica utilizada en el algoritmo de aprendizaje es la del gradiente descendente con un término de momento.

En la fase de entrenamiento la red aprenderá a clasificar a los alumnos en el grupo de éxito o de fracaso, atendiendo a las características que hemos mencionado. Esta fase parará cuando se alcance el número de épocas prefijadas o bien el error cuadrático medio en el conjunto de validación no mejore durante 50 épocas consecutivas, con objeto de evitar el sobreaprendizaje.

A continuación presentamos las denominadas matrices de confusión, que recogen los porcentajes de éxitos y fracasos clasificados correctamente e incorrectamente por la correspondiente red. Así, en la tabla 1 podemos ver que el 55'36% de los alumnos de Empresariales en el conjunto de entrenamiento que realmente tienen éxito son clasificados correctamente por la red, y el 75'31% de los que fracasan también. Análogamente, en la tabla 2 se recogen los mismos resultados para el conjunto de datos destinados a testar. Al ser ambas matrices muy similares el poder de generalización de la red se considera bueno.

<i>Diplomatura de Ciencias Empresariales</i>		
	ÉXITO	FRACASO
ÉXITO	55'36%	44'64%
FRACASO	24'69%	75'31%

TABLA 1.- Matriz de confusión para los datos de entrenamiento de la Diplomatura de Ciencias Empresariales

<i>Diplomatura de Ciencias Empresariales</i>		
	ÉXITO	FRACASO
ÉXITO	56'67%	43'33%
FRACASO	27'08%	72'92%

TABLA 2.- Matriz de confusión para los datos de test de la Diplomatura de Ciencias Empresariales

Los resultados para los alumnos de las licenciaturas se presentan en las tablas 3 y 4, en las que se observa una mejor clasificación del fracaso que en el caso de la red obtenida para la diplomatura. Como en la situación anterior, el poder de generalización es también aceptable.

<i>Licenciaturas de Economía y ADE</i>		
	ÉXITO	FRACASO
ÉXITO	58'67%	41'33%
FRACASO	7'14%	92'86%

TABLA 3.- Matriz de confusión para los datos de entrenamiento de las licenciaturas de Economía y ADE

<i>Licenciaturas de Economía y ADE</i>		
	ÉXITO	FRACASO
ÉXITO	42'85%	57'15%
FRACASO	7'35%	92'65%

TABLA 4.- Matriz de confusión para los datos de test de las licenciaturas de Economía y ADE

4. CONCLUSIONES

Desde el punto de vista metodológico, la principal aportación de este trabajo es la utilización de una nueva técnica de clasificación con la que se suelen obtener mejores resultados predictivos que con los métodos tradicionales (análisis discriminante y regresión logística). Esta técnica nos permite detectar *a priori* los alumnos con menos posibilidades de éxito, de modo que llevando a cabo un cierto seguimiento y/o apoyo se podrían mejorar sus resultados finales.

De todos modos, puesto que el porcentaje de alumnos mal clasificados es bastante elevado en algunos casos, sería interesante ampliar el trabajo considerando nuevas variables input que recojan información acerca del comportamiento del alumno durante cierto periodo de tiempo inmediatamente posterior al inicio de los estudios.

Finalmente, esta técnica puede trasladarse a otros contextos interesantes en el campo de la educación no universitaria, como a la predicción del éxito en las pruebas de acceso a la universidad.

5. REFERENCIAS BIBLIOGRÁFICAS

- BROCKETT, PL., COOPER, WW., GOLDEN, LL., XIA, X. (1997): "A case study in applying networks to predicting insolvency for property and casualty insurers". *Journal of the Operational Research Society*, Vol. 48, pp. 1153-1162.
- CURRAM, S.P., MINGERS, J. (1994): "Neural networks, decision tree induction and discriminant analysis: an empirical comparison". *Journal of the Operational Research Society*, Vol. 45, No. 4, pp. 440-450.
- DESAI, V.S., CROOK, J. N., OVERSTREET, G.A. (1996) "A comparison of neural networks and linear scoring models in the credit union environment". *European Journal of Operational Research*, Vol. 95, pp. 24-37.
- KURKOVA, V. (1992): "Kolmogorov theorem and multilayer neural networks". *IEEE.ASSP Magazine*, Vol. 1, pp. 4-22.
- QUESADA, V. (2000): "Predicción dinámica mediante redes neuronales". *Perspectivas en estadística e investigación operativa*. (Pascual, A. y Parras, L., eds.) Colección Techné, Universidad de Jaén.
- RUMERLHART, D.E., HINTON, G., WILLIAMS, R. (1986): "Learning representation by back-propagating errors". *Nature*, Vol. 323, No. 9, pp. 533-536.
- SANTIN, D. (1999): "Detección de alumnos de riesgo y medición de la eficiencia de centros escolares mediante redes neuronales". *Documento de Trabajo 9902*, Campus de Somosaguas. UCM.
- TAM, K.Y., KIANG, M.Y. (1992): "Managerial applications of neural networks: The case of bank failure predictions". *Management Science*, Vol. 38, no. 7, pp. 926-947.
- WANG, J., MALAKOOTI, B. (1993): "Characterization of training errors in supervised learning using gradient-based rules". *Neural Networks*, Vol. 6, pp. 1073-87.

