La supervivencia de Haberman mediante Redes de Neuronas Artificiales

Pérez Fraguela, Eduardo

eduardo.perez.fraguela@udc.es

Facultade de Informática, Universidade da Coruña, A Coruña, Spain.

INTRODUCCIÓN

La base de datos de la supervivencia de Haberman, o Haberman's Survival, contiene casos de un estudio que se realizó entre los años 1958 y 1970, en el Hospital Billings de la Universidad de Chicago. Este estudio trata sobre la supervivencia de pacientes que se habían sometido a cirugía por cáncer de mama. En esta base de datos se trataron tres parámetros: la edad del paciente al momento de la operación, el año de la operación y el número de ganglios axilares positivos detectados. Estos datos se trataron en pacientes, de los cuales la base de datos contiene trescientos seis. El objetivo es, dados esos tres parámetros, determinar el estado de supervivencia del paciente tras la operación: el paciente sobrevivió cinco años o más, o el paciente falleció dentro de los cinco años. Es, por tanto, un problema de clasificación.

Los tres parámetros de entrada son valores numéricos (enteros), con los que no ha sido necesario realizar ninguna codificación. Al contrario, la salida ha tenido que ser codificada con un booleano. La siguiente tabla muestra los valores de salidas deseadas utilizadas:

Supervivencia	Salidas deseadas
Sobrevivió cinco años o más	1
Falleció dentro de los cinco años	0

Tabla 1. Codificación usada para la clase de salida.

EXPERIMENTOS REALIZADOS

Para resolver este problema se han utilizado Redes de Neuronas Artificiales. En concreto, se han probado distintas arquitecturas: [ ], [3], [5], [10], [15], [5 3], [4 3], [3 5], [10 5]. Como podemos observar en cuanto a las capas, hemos empleado tanto una como dos capas ocultas. Los elementos con una capa oculta son aquellos formados por 3, 5, 10 y 15 neuronas. Mientras que los que poseen dos capas ocultas son el [5 3], [4 3], [3 5], [10 5]. Estas capas se han entrenado 50 veces cada una.

La siguiente tabla muestra la arquitectura empleada, así como los resultados en test obtenidos para cada una de ellas.

Arquitectura	Media de resultados en test	Desviación típica de resultados en test
[ ]	74.74%	5.57
[3]	73.61%	8.22
[5]	72.61%	6.35
[10]	74.83%	6.69
[15]	74.78%	5.43
[5 3]	73.52%	6.90
[4 3]	73.22%	5.90
[3 5]	72.78%	5.93
[10 5]	72.52%	5.67

Tabla 2. Resultados obtenidos en test para cada arquitectura.

CONCLUSIONES

Observando los datos, podemos concluir que los resultados son mejorables para el problema. Tienen un porcentaje reducido de acierto para el caso, no alcanza ni el **90%**. Estos datos se podrían mejorar con una mejor base de datos. Las arquitecturas más o menos se mantienen sobre la misma línea, entre el **72%** y el **75%** de media, pero la que mayor porcentaje posee es la arquitectura de [10], con un porcentaje de **74.83%**. A diferencia de la de [10], la de [10 5] es la que menor porcentaje posee con un **72.52%**. El mayor inconveniente que tuve al resolver el problema fue encontrar una base de datos que me atrajese. Por lo demás, creo que es un problema sencillo de resolver y de entender. En mi opinión, resulta viable usar Redes de Neuronas Artificiales debido a que ofrece buenos resultados para la escasez de datos, crea modelos generales, tiene un gran ratio de predicción y es muy configurable. Sería interesante tener más datos para el empleo de Deep Learning.

REFERENCIAS BIBLIOGRÁFICAS

Haberman, S. J. (1976). Generalized Residuals for Log-Linear Models, Proceedings of the 9th International Biometrics Conference, Boston, pp. 104-122.

Landwehr, J. M., Pregibon, D., and Shoemaker, A. C. (1984), Graphical Models for Assessing Logistic Regression Models (with discussion), Journal of the American Statistical Association 79: 61-83.

Lo, W.-D. (1993). Logistic Regression Trees, PhD thesis, Department of Statistics, University of Wisconsin, Madison, WI.