Un Modelo para la Predicción de Recidiva de Pacientes Operados de Cáncer de Mama (CMO) Basado en Redes Neuronales.

J.A. Gómez Ruiz[(1)], J.M. Jerez Aragonés[(1)], J. Muñoz Pérez[(1)], E. Alba Conejo[(2)]

[(1)]Dpto. Lenguajes y Ciencias de la Computación
Universidad de Málaga.
Campus de Teatinos s/n. 29071 Málaga
[(2)]Servicio de Oncología del Hospital Clínico Universitario.
29071 Málaga.
{janto,jja}@lcc.uma.es

Resumen

La predicción de recidiva en pacientes que han sido operados de cáncer de mama juega un papel muy importante en tareas médicas como el diagnostico y la planificación del tratamiento que hay que realizarle al mismo. En la actualidad, los expertos médicos están llevando a cabo estas tareas usando técnicas no numéricas. Las redes neuronales artificiales se muestran como una herramienta potente para el análisis de conjuntos de datos donde hay relaciones no lineales entre los datos a estudio y la información a ser predecida. En este artículo estimamos tanto la probabilidad de clasificación correcta como la regla de Bayes utilizando un perceptrón multicapa, que nos permite, al mismo tiempo, conocer la precisión de la regla de decisión obtenida. Este estudio se ha aplicado en la predicción de recidiva de pacientes operados de cáncer de mama, usando para ello datos clínico-patológicos (tamaño del tumor, edad del paciente, receptores de estrógenos, etc.) procedentes del servicio médico de Oncología del Hospital Clínico Universitario de Málaga. Se han estudiado diferentes topologías del perceptrón multicapa para obtener la mejor precisión en la predicción. Los resultados actuales muestran que, después del proceso de aprendizaje, el modelo teórico final propuesto es apropiado para hacer predicciones de la probabilidad de recidiva en diferentes intervalos de tiempo.

Palabras clave: Perceptrón, Regla de Bayes, Clasificación, Diagnóstico Médico, Predicción de Recidiva, Probabilidad de Clasificación Correcta.

1. Introducción

La predicción es un intento de diagnosticar con precisión la evolución de un sistema específico usando para ello la información obtenida a partir de un conjunto concreto de variables que describen dicho sistema.

El problema que se plantea frecuentemente en medicina clínica es como llegar a una conclusión sobre el pronóstico de pacientes cuando se presentan con una información clínica compleja. Los expertos clínicos usualmente toman decisiones basadas en una simple dicotomización de variables en clasificaciones favorables y desfavorables (McGuire, 1990). En este trabajo, analizamos el proceso de decisión que se presenta cuando los pacientes con un cáncer de mama primario reciben una cierta terapia para eliminarlo. En este punto es muy importante estimar la probabilidad de que el paciente sufra una recaída en su enfermedad de manera que el riesgo y los beneficios esperados de terapias específicas se puedan comparar.

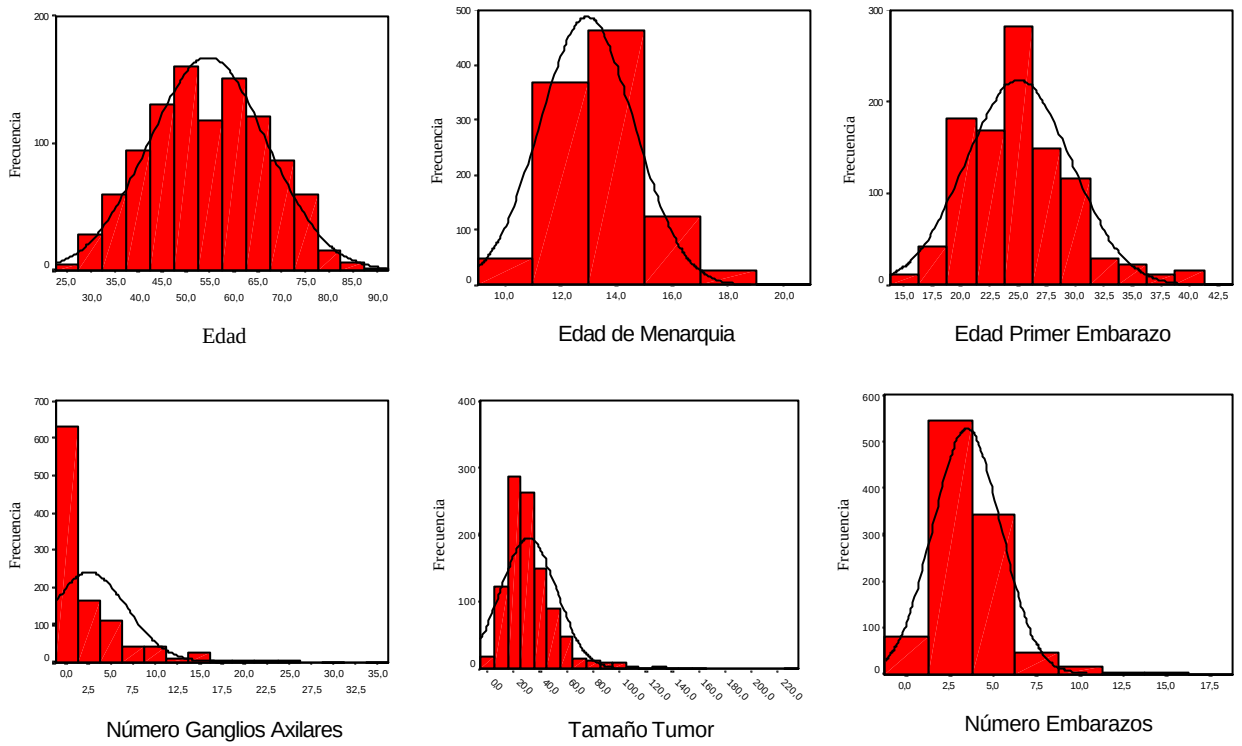

Figura 1. Histogramas de los factores de pronóstico

	Mínimo	Máximo	Rango	Media	Desv. Típica
Edad	24	89	65	54,59	12,27
Edad de menarquia	9	20	11	12,94	1,69
Edad primer embarazo	15	42	27	25,04	4,59
Nº ganglios axilares	0	34	34	2,52	4,29
Tamaño del tumor	0	230	230	31,15	21,03
Número embarazos	1	18	17	3,48	1,96

Tabla 1. Resumen de los datos de los pacientes: media, desviación típica y rangos

Las redes neuronales se aplican en un amplio rango de problemas (Gorman, 1988; O'Neill, 1991; Qian, 1988) en los que en múltiples casos superan en resultados a los modelos estadísticos clásicos (White, 1989). Baxt (1990) mostró la exactitud predictiva de los modelos de redes neuronales artificiales en el diagnóstico médico. En este caso, nosotros utilizaremos la capacidad de las redes neuronales para reconocer las relaciones complejas, y altamente no lineales, que se presentan a la hora de realizar el diagnóstico médico.

Algunos autores (Ravdin, 1992; Jefferson, 1996) han modelado sistemas para la predicción de recidiva de pacientes después de haber sido operados de cáncer de pulmón y de mama. Ellos hacen uso de las redes neuronales para llevar a cabo análisis de supervivencia junto con diferentes estimadores de supervivencia que manejan datos censurados de pacientes. Esto implica que los factores de pronóstico, por ejemplo en cánceres de mama con tratamiento adyuvante después de la cirugía, sean independientes del tiempo transcurrido, pero ésto no es realmente cierto. Es decir, que la influencia del factor de pronóstico no sea la misma para diferentes intervalos de tiempo. Diferentes técnicas para la estimación de la supervivencia, como el análisis de Kapplan-Meier (Kaplan et al., 1958) y el modelo de regresión de Cox (Cox, 1972), suponen que la influencia de un factor de pronóstico no cambia durante el tiempo. Además, al existir un "pico" de recurrencia en la

distribución de probabilidad de recidiva (Alba et al., 1999) se demuestra que dicha probabilidad no es la misma a lo largo del tiempo, dependiendo por tanto del periodo en el que se encuentre el paciente.

Debido a todo esto, nosotros proponemos un sistema basado en redes neuronales con topologías específicas para cada intervalo de tiempo durante el periodo de seguimiento de los pacientes.

Este artículo está organizado de la siguiente forma: en la sección 2 presentamos el material experimental usado; en la sección 3 presentamos la regla de decisión, el modelo de pronóstico y los resultados obtenidos; y finalmente, en la sección 4 exponemos las conclusiones y el trabajo futuro.

2. Material experimental

Los datos de los pacientes usados en el análisis proceden de una base de datos del Servicio Médico de Oncología del Hospital Clínico Universitario de Málaga. Está base de datos contiene un total de 1035 registros correspondiente cada uno de ellos a un paciente y han sido recopilados a lo largo de 30 años. Cada registro está estructurado en 85 campos que contienen información acerca de medidas postquirúrgicas, datos personales, tipo de tratamiento, edad, etc. Después de consultar a los expertos médicos, el conjunto de variables que se han considerado más oportunas para preparar el modelo predictivo consta de los siguientes factores de pronóstico: edad del paciente, tamaño del tumor, número de ganglios axilares, número de embarazos, edad del primer embarazo y la edad de menarquía (primera menstruación). En la tabla 1 mostramos la media, máximo, mínimo, rango y desviación típica de todas las variables que representan a dichos factores de pronóstico. Los histogramas correspondientes a dichos factores de pronóstico los mostramos en la figura 1.

3. El Sistema de Diagnóstico

3.1. Aproximando la Regla de Decisión de Bayes

El problema que se nos presenta es el siguiente: dado un paciente determinado que presenta unos factores de pronóstico concretos ¿sufrirá recaída de la enfermedad durante los intervalos del periodo de seguimiento? Para responder a esta pregunta necesitamos una regla de decisión. El criterio que vamos a seguir para elegir dicha regla va a ser el de maximizar la probabilidad de clasificación correcta, es decir vamos a tratar de determinar la regla de decisión de Bayes (Duda et al., 1973) que viene dada por la expresión:

$$\boldsymbol{f}_i(x) = \begin{cases} 1 & \text{si } p(C_i / x) \geq p(C_k / x) \ \forall k \\ 0 & \text{en otro caso} \end{cases}$$

donde $p(C_i/x)$ es la distribución de probabilidad *a posteriori* y $\boldsymbol{f}_i$ es la probabilidad de clasificar el patrón x en la clase C_i.

Por ello necesitamos determinar la probabilidad de clasificación correcta de Bayes en nuestro problema. Sea $p(C_i)$ la probabilidad *a priori* de la clase C_i, donde $i = 1$ identifica a la clase "recidivar" e $i = 2$ identifica a la clase "no recidivar", y sea p_{ii} la probabilidad condicionada de clasificar un patrón de la clase C_i en C_i, entonces la probabilidad de clasificación correcta viene dada por la expresión:

$$\begin{aligned} p &= p(C_1)\cdot p_{11} + p(C_2)\cdot p_{22} \\ &= p(C_1)\cdot \int_{\Re^N} \boldsymbol{f}_1(x)\cdot p(x/C_1)\,dx + \\ &\qquad p(C_2)\cdot \int_{\Re^N} \boldsymbol{f}_2(x)\cdot p(x/C_2)\,dx \\ &= \int_{A=\{x:\, p(C_1/x)\geq p(C_2/x)\}} p(C_1)\cdot p(x/C_1)\,dx + \\ &\qquad \int_{\bar{A}} p(C_2)\cdot p(x/C_2)\,dx \\ &= \int_{\Re^N} p(C_1)\cdot p(x/C_1)\,dx + \\ &\qquad \int_{\bar{A}} \big(p(C_2)\cdot p(x/C_2) - p(C_1)\cdot p(x/C_1)\big)\,dx \\ &= p(C_1) + \int_{\bar{A}} \big[p(C_2/x) - p(C_1/x)\big]\cdot p(x)\,dx \\ &= p(C_1) + \int_{\bar{A}} \big(1 - 2\,p(C_1/x)\big)\cdot p(x)\,dx \end{aligned} \qquad (1)$$

De la misma forma tenemos que

$$p = p(C_2) + \int_{A} \big(2\,p(C_1/x) - 1\big)\cdot p(x)\,dx \qquad (2)$$

De las ecuaciones (1) y (2) obtenemos

$$p = \frac{1}{2} + \int_{\Re^N} \left| p(C_1/x) - \frac{1}{2} \right| \cdot p(x)\,dx \qquad (3)$$

Por tanto $p \geq Max\{p(C_1), p(C_2)\}$ que nos da una idea de cual es el valor más pequeño que podemos conseguir para la probabilidad de clasificación correcta.

La distribución de probabilidad *a posteriori*, $p(C_i/x)$, es desconocida en el problema que se nos presenta, por lo que tenemos que estimarla y obtener así una probabilidad de clasificación correcta aproximada. Funahashi K. (1998) demuestra que en una red neuronal de tres capas, usando un algoritmo de retropropagación donde, para el entrenamiento, se asigna salida uno, cuando el patrón de entrada pertenece a la clase C_1, y salida cero, cuando pertenece a la clase C_2, la salida de la red tiende a la distribución de probabilidad *a posteriori* $p(C_1/x)$. Es decir, al finalizar el proceso de aprendizaje, tenemos que

$$p(C_1 / x) \cong F(x, t, w) \tag{4}$$

donde $F(x,t,w)$ es la salida de la red para un patrón de entrada x dado, y siendo t y w las matrices de pesos sinápticos obtenidas tras el proceso de aprendizaje. Por lo tanto, la regla de decisión de Bayes estimada viene dada por la expresión

$$\boldsymbol{f}(x) = \begin{cases} 1 & \text{si } F(x,t,w) \geq 1/2 \\ 0 & \text{si } F(x,t,w) < 1/2 \end{cases} \tag{5}$$

que nos da la probabilidad de clasificación del patrón x en la clase C_1. Así, si $\boldsymbol{f}(x) = 1$, el patrón se clasifica en la clase C_1 y si $\boldsymbol{f}(x) = 0$, se clasifica en la clase C_2.

De las ecuaciones (3) y (4) obtenemos que la probabilidad estimada de clasificación correcta de Bayes viene dada por la expresión

$$\hat{p} = \frac{1}{2} + \frac{1}{n}\sum_{i=1}^{n}\left|F(x_i,t,w) - \frac{1}{2}\right| \tag{6}$$

donde n es el número total de pacientes en estudio.

Las probabilidades p_{11} y p_{22} se pueden también estimar mediante la red neuronal multicapa como

$$\hat{p}_{11} = \frac{1}{m} \cdot \sum_{\{x \in C_1:\ F(x,t,w) \geq 1/2\}} F(x,t,w)$$

$$\hat{p}_{22} = \frac{1}{n-m} \cdot \sum_{\{x \in C_2 : F(x,t,w) \leq 1/2\}} \left(1 - F(x,t,w)\right)$$

donde m es el número total de pacientes que recidivan.

La probabilidad estimada de clasificación correcta de Bayes dada en la ecuación (6) es la cota superior que podemos alcanzar en nuestro problema, es decir, nos da una idea de la dificultad de la clasificación en donde la mejor regla de decisión nos daría a lo sumo dicha probabilidad.

3.2. El Modelo Propuesto

Los factores de pronóstico utilizados en el cáncer de mama operable, cuando se usa terapia adyuvante después de la cirugía, son dependientes del periodo de tiempo en estudio. Esto quiere decir que la importancia de un factor de pronóstico no es la misma para los diez primeros meses que, por ejemplo, para el intervalo comprendido entre los cincuenta y sesenta meses. En diferentes técnicas para la estimación de supervivencia, como el análisis de Kapplan-Meier (Kaplan et al., 1958) y el modelo de regresión de Cox (Cox, 1972), se supone que la importancia del factor de pronóstico no cambia durante la evolución del tiempo y esto no es cierto en nuestro caso. Hay que añadir también que la probabilidad de recidiva del paciente no es la misma a lo largo del tiempo, ya que existe un "pico" de recurrencia en la distribución de probabilidad de recidiva que ha sido demostrado empíricamente (Alba et al., 1999).

Considerando todo esto y la justificación de la regla de decisión propuesta en la ecuación (5), proponemos un esquema basado en diferentes topologías de redes neuronales, específicas para cada intervalo de tiempo en los que se ha dividido el periodo de tratamiento de los pacientes. Este esquema consta de un sistema de perceptrones multicapa y de una unidad de disparo que implementa dicha regla de decisión (ver la figura 2). El sistema neuronal computa un conjunto de atributos extraídos del registro del paciente y obtiene como salida una estimación de la probabilidad *a posteriori* de recidiva para dicho paciente. La unidad de disparo recoge la salida del sistema neuronal y nos da el diagnóstico atendiendo a la regla de decisión propuesta en la ecuación (5).

Todas los sistemas neuronales, considerados para cada intervalo de tiempo, tienen tres capas (entrada, oculta y salida) y usan la tangente hiperbólica como función de transferencia en la capa oculta, y la función logística en la capa de salida.

Un aspecto crucial para poder realizar aprendizaje y diagnóstico en la red neuronal es seleccionar dos conjuntos independientes de datos procedentes de la base de datos de los pacientes, que serán usados respectivamente para el entrenamiento de la red y para validar la eficacia de la predicción (Haykin, 1994).

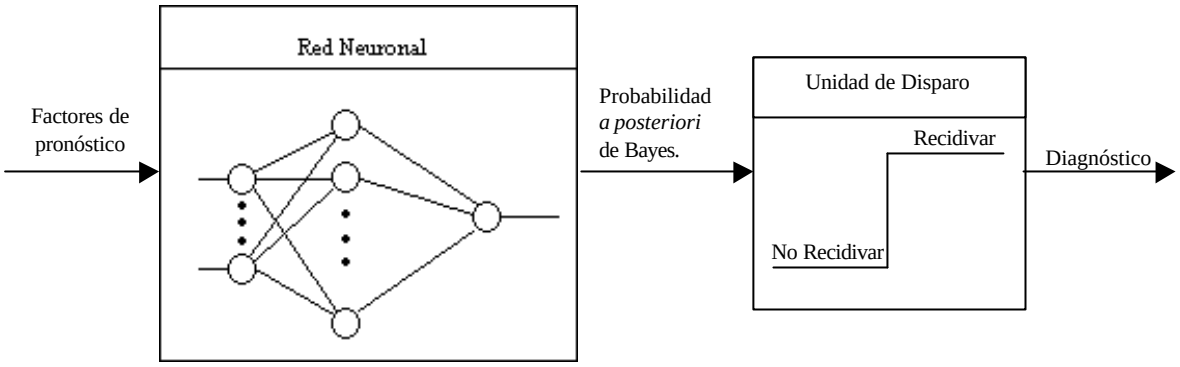

Figura 2. Sistema de diagnóstico propuesto

Intervalos de tiempo (Nº meses)	Número de Pacientes	Probabilidad *a priori* de recidivar
I_1 (0 – 10)	845	6,75 %
I_2 (10 – 20)	741	10,66 %
I_3 (20 – 30)	681	6,9 %
I_4 (30 – 40)	600	6,33 %
I_5 (40 – 50)	520	4,36 %
I_6 (50 – 60)	466	5,36 %
I_7 (> 60)	466	7,08 %

Tabla 2. Número de pacientes y probabilidad *a priori* para cada intervalo de tiempo

Durante la fase de entrenamiento, los factores de pronóstico se introducen periódicamente y en la misma proporción hasta que la red nos proporcione como salida un uno cuando el estado del paciente sea "recidivar" y cero cuando el estado sea "no recidivar". Los valores de los pesos sinápticos de la red se van actualizando mediante el algoritmo de retropropagación de Levenberg-Marquardt (Patterson, 1996).

Previamente, para poder facilitarle los datos a la red, hemos tenido que realizar preprocesamiento de los factores de pronóstico. Primero hay que estudiar todos los factores de pronóstico y su distribución para eliminar los valores perdidos y reducir el impacto de los que se encuentran en la cola de las distribuciones; y segundo, normalizar todos los factores de pronóstico para que se extiendan dentro del rango central de la función de transferencia de la capa oculta de la red ([-1,1] para la tangente hiperbólica).

Los subconjuntos de datos correspondientes a cada intervalo de tiempo estudiado se han seleccionado de los 1035 pacientes de la base de datos del servicio de oncología y se han clasificado en las clases C_1 y C_2. Dicha clasificación se hace para cada intervalo teniendo en cuenta todos los pacientes de la base de datos, de forma que, dado un intervalo de tiempo I_i en estudio (ver tabla 2 para los intervalos), los pacientes seleccionados, y la clase a la que se asignan, se obtienen según las siguientes reglas:

1. Pacientes del intervalo I_i: se contabilizan como de la clase C_1 aquellos pacientes cuyo estado de supervivencia para dicho intervalo sea recidivado. El resto se ignoran.
2. Pacientes del intervalo I_j (j<i): se contabilizan para la clase C_2 aquellos pacientes cuyo estado de supervivencia era recidivado. El resto se ignoran.
3. Pacientes del intervalo I_k (i<k): Se contabilizan todos para la clase C_2.

En la tabla 2 se muestra, para cada intervalo de tiempo, tanto el número total de pacientes seleccionados, como la probabilidad *a priori* de recidivar, que no es sino el tanto por ciento de pacientes de la clase C_1 con respecto al total para ese intervalo.

La inicialización de los pesos es importante en el proceso de aprendizaje con redes neuronales artificiales. Con el fin de mejorar el error cuadrático al final del proceso de aprendizaje, se han realizado diferentes inicializaciones de pesos para obtener el mejor conjunto de valores de los pesos sinápticos.

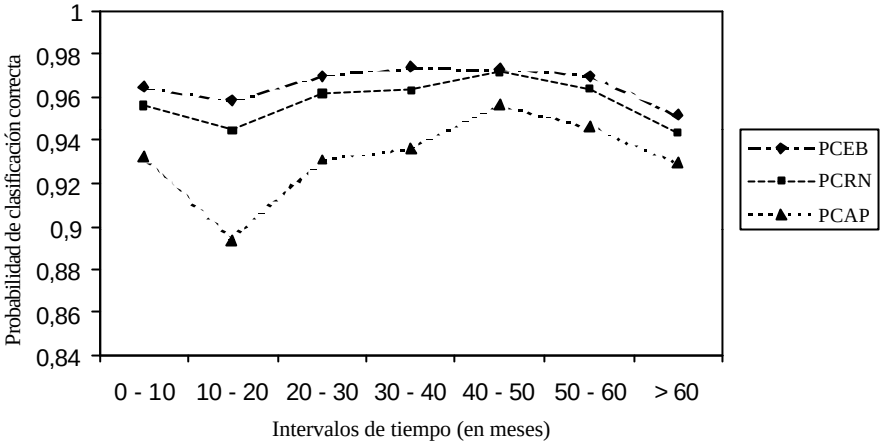

Figura 3. Probabilidad de clasificación correcta estimadas para los diferentes intervalos de tiempo según que hayamos utilizado la regla de Bayes (PCEB), la regla de Bayes obtenida por la red neuronal (PCRN) y las probabilidades *a priori* (PCAP).

Además, para mejorar la precisión del modelo se ha usado la técnica *cross-validation* (Janssen et al., 1988) para obtener el valor medio de las probabilidades de clasificación correcta.

3.3 Resultados

Para poder evaluar la probabilidad de clasificación correcta de nuestro sistema, hemos dibujado simultáneamente en la figura 3, y para todos los periodos de tiempo, la probabilidad de clasificación correcta estimada cuando se utiliza la regla de Bayes (PCEB), la probabilidad de clasificación correcta estimada cuando se utiliza la red neuronal (PCRN) y la probabilidad *a priori* de no recidiva de los pacientes (PCAP).

En dicha figura podemos observar varios resultados. Primero, el sistema propuesto (PCRN) siempre mejora la probabilidad *a priori* (PCAP) de no recidivar. Esto es importante destacarlo dada la dificultad del problema, ya que los valores para la PCAP son muy altos en cada intervalo de tiempo. Además, esta mejora es mayor en el intervalo más crítico (I_2) del periodo de recuperación del paciente (Alba E. et al., 1999). Segundo, la PCRN es siempre menor que la PCEB y sigue la ascendencia de la misma, como era de esperar. Además, es importante apuntar que la diferencia entre ambas (PCRN y PCEB) no es significativa, lo que significa que la regla propuesta en la ecuación (5) es un buen estimador de la regla de decisión de Bayes.

4. Conclusiones y Trabajo Futuro

Se han utilizado diferentes topologías de perceptrones multicapa para obtener la mejor precisión en la probabilidad de clasificación correcta de pacientes que recidivan después de haber sido operados de cáncer de mama, usando para ello datos clínico-patológicos. El sistema final propuesto, basado en una estimación de la regla de decisión de Bayes mediante redes neuronales, consigue predicciones sobre la probabilidad que tiene el paciente de recidivar en diferentes intervalos de tiempo durante su periodo de tratamiento con una tasa de error muy pequeña. La tasa de error se ha podido comprobar al conseguir estimar la probabilidad de clasificación correcta (cuando se utiliza la regla de Bayes verdadera en lugar de la estimada) mediante la salida suministrada por la red neuronal.

El siguiente paso dentro de esta línea de trabajo es intentar buscar qué tipo de relación hay entre el valor de los pesos de la red neuronal y los factores médicos de diagnóstico usados, ya que uno de los problemas que tienen las redes neuronales es poder entender como procesan la información y cuál es la función no lineal que se constituye tras la matriz de pesos y las funciones de transferencia. Estamos actualmente trabajando en este problema mediante dos caminos distintos. El primero es entrenar la red neuronal alternativamente con diferentes entradas, incluyendo y omitiendo pares de variables para diferentes simulaciones. Es importante destacar que

el registro de cada paciente del hospital tiene un total de 85 campos conteniendo distinta información de otros muchos posibles factores de pronóstico, además de los investigados en este trabajo. El segundo enfoque es introducir una metodología basada en algoritmos genéticos para la inducción automática de las topologías de redes neuronales, para identificar cuáles son aquellas variables que tienen más importancia en la mejora del pronóstico.

Agradecimientos

Agradecemos a Ana Sánchez y al personal del Servicio de Oncología del Hospital Clínico Universitario de Málaga tanto sus comentarios como su ayuda para la realización de este trabajo.

Referencias

Alba E. et. al. Estructura del patrón de recurrencia en el cancer de mama operable (CMO) tras el tratamiento primario. Implicaciones acerca del conocimiento de la historia natural de la enfermedad. 7º Congreso de la Sociedad Española de Oncología Médica, Sitges, Barcelona. Abril 1999.

Baxt WG. Application of neural networks to clinicla medicine. Lancet, 346:1135-8, 1995.

Cox DR: Regression models and life tables. J R Stat Soc [B] 34:187-220, 1972.

Duda RO, Hart PE: Pattern classification and scene analysis. New-York: John Wiley and Sons, 1973.

Funahashi K: Multilayer neural networks and Bayes decision theory. Neural Networks, 11:209-213, 1998.

Gorman RP, Sejnowski TJ: Analysis of hidden units in a layered network trained to classify Sonar targets. Neural Networks 1:75-89, 1988.

Haykin S.: Neural Networks, Theory and Apllications. Macmillan College Publishing Company, 1994.

Janssen P. et. al. Model Structure Selection for multivariable Systems by Cross-Validation. International Journal of Control, 47:1737-1758, 1988.

Jefferson M, Pendleton N, Lucas B, Horan M: Comparison of a genetic algorithm neural network with logistic regression for predicting outcome after surgery for patients with nonsmall cell lung carcinoma. American Cancer Society, 1996.

Kaplan SA, Meier, P: Nonparametric estimation from incomplete observations. J. Am. Stat. Assoc., 53, 457-481, 1958.

McGuire WL, Tandom AT, Allred DC, Chamnes GC, Clark GM: How to use prognostic factors in axillary node-negative breast cancer patients. J Natl Cancer Inst 82:1006-1015, 1990.

Patterson, D.W. : Artificial Neural Netwoks, Theory and Applications. Prentice Hall, 1996.

O´Neill Mc: Training back-propagation neural networks to define and detect DNA-binding sites. Nucleic Acids Res 19:313-318, 1991.

Qian N, Sejnowski TJ: Predicting the secondary structure of globular proteins using neural network models. J Mol Biol 202:865-884, 1988.

Ravdin, PM, Clark, GM: A practical application of neural network analysis for predicting outcome of individual breast cancer patients. Breast Cancer Research and Treatment, 22: 285-293, 1992.

White H: Learning in artificial neural networks: a statistical approach. Neural Computation 1:425-464, 1989.