UNIVERSIDAD CARLOS III DE MADRID

CAMPUS DE TOLEDO

MASTER EN CC. ACTUARIALES Y FINANCIERAS

TRABAJO FIN DE MASTER

SISTEMA PREDICTIVO DE TABLAS DE MORTALIDAD MEDIANTE REDES DE NEURONAS Y ALGORITMOS GENÉTICOS

Autor: Juan David Arias Rodríguez

Tutores: José Miguel Rodríguez-Pardo del Castillo
Jesús Ramón Simón del Potro

30 DE JUNIO, 2017

En caso de obtener una calificación igual o superior a 8.0 Notable, autorizo la publicación de este trabajo en el centro de Documentación de la Fundación Mapfre.

X	Sí, autorizo a su publicación.
	No, desestimo su publicación.

Fdo.

A mi familia,Yolanda,
Arane y Enara.

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

RESUMEN

El presente trabajo, tiene como objetivo la predicción de las tablas de mortalidad, por medio del uso de una red de neuronas en base radial, apoyándose en algoritmos genéticos, para encontrar la mejor arquitectura de la red de neuronas.

El riesgo de longevidad es uno de los sub-riesgos encuadrados dentro del riesgo de suscripción de vida, siendo el sub-riesgo de mortalidad otro de esos riesgos. Para su cálculo, la técnica actuarial se basa en las llamadas tablas de mortalidad, por eso es muy importante el poder contar con algún método, que nos pueda dar una aproximación a futuro de cómo podrían ir variando estas tablas a lo largo del tiempo.

A tal efecto, se han implementado desde cero tanto el funcionamiento de la red de neuronas, como la del algoritmo genético sin depender de librerías de terceros. El lenguaje escogido ha sido C/C++, que es un lenguaje compilado que permite una gran velocidad de cómputo.

ABSTRACT

The present work aims to predict the mortality tables by using a radial basis neural network, finding the best architecture of the network of neurons by means of genetic algorithms.

The longevity risk is one of the sub-risks framed within the life underwriting risk, with the under-risk of mortality being another of those risks. For its calculation, the actuarial technique is based on mortality tables, so it is very important to have some method, which can give us a future approximation of how these tables could vary over time.

To that effectd, both the operation of the network of neurons and the genetic algorithm have been implemented from scratch without depending on third-party libraries. The chosen language has been C/C ++, which is a compiled language that allows a high computational speed.

CAPITULO 1 – INTRODUCCIÓN

1. INTRODUCCIÓN

1.1. MOTIVACIÓN Y OBJETIVOS DEL PROYECTO

1.2. CONTENIDO DE LOS CAPÍTULOS

CAPÍTULO 1 - INTRODUCCIÓN

En este primer capítulo se exponen las ideas generales que han llevado a la realización de este trabajo fin de máster, así como un resumen del contenido de cada uno de los capítulos del proyecto.

1.1 - Motivación y objetivos del TFM

Para la práctica actuarial, es de suma importancia el poder calcular unas provisiones matemáticas, que permitan atender las obligaciones contraídas. Si pensamos por ejemplo en el típico caso de una renta vitalicia a partir de la edad de jubilación, entenderemos que si los cálculos en cuanto a mortalidad están errados, y los tomadores empiezan a vivir más, esto generará unos gastos que podrían llegar a ser inadmisibles para aquellos que contrajeron esas obligaciones.

Atendiendo a la longevidad, y más allá del aumento del cáncer, diabetes, enfermedades del aparato circulatorio, etc, se ha podido comprobar que la esperanza de vida en los países industrializados aumenta a una tasa estimada de 15 minutos por hora. Esto provoca que los cálculos basados en tablas de mortalidad desfasadas den como resultado provisiones subestimadas.

Existen varios métodos para intentar pronosticar la tendencia de la longevidad, de los más utilizados se encuentran los modelos paramétricos como Lee Carter, P-Spline, Renshaw-Habermann, Kannistö, etc. En este trabajo se va a aplicar un algoritmo de machine learning, de hecho dos algoritmos, el principal basado en redes de neuronas de base radial, mientras que otro de apoyo al primero es un algoritmo genético.

Estos algoritmos como son denominados dentro de la ingeniería informática, o ciencias de la computación, pertenecientes más concretamente a la especialidad de Inteligencia Artificial, son de aplicación a la resolución de múltiples problemas en muchos ámbitos de las ciencias. Como sus nombres nos indican provienen del estudio del cerebro en el

caso de las redes neuronales artificiales (RNA), y de la teoría evolutiva en el caso de los algoritmos genéticos (AG).

Las redes de neuronas, funcionan como sistemas de clasificación, o como métodos de interpolación o aproximación universal de funciones, y se han aplicado con éxito a campos tan variopintos como, la medicina, la economía, la predicción de mareas, del tiempo, de precios, reconocimiento de voz, reconocimiento de imágenes y un largo etcétera.

Por su parte los algoritmos genéticos son algoritmos de búsqueda, basados en la selección natural y la genética, generalmente usados en problemas de optimización, donde presentan su mayor campo de actuación. En particular se usan para la resolución de aquellos problemas para los que no existen técnicas concretas de resolución, o existiendo éstas no son óptimas.

Las RNA en general, y las de base radial en particular, cuentan con una base matemática bastante compleja, así como con una serie de parametrizaciones, como el número de neuronas de cada una de las capas, el tipo de aprendizaje, sus funciones de activación, sus coordenadas etc, que hacen que la estimación de estos parámetros sea más un trabajo de artesanía, que se va depurando mediante prueba y error.

Dada la dificultad añadida de encontrar buenas parametrizaciones para una RNA, es donde entra en juego el uso de los algoritmos genéticos, de cuyos resultados podremos estimar aquel rango de parámetros que funcionan bien con el problema que queremos resolver, es decir el del pronóstico de los valores de una tabla de mortalidad.

1.2- Contenido de los capítulos

A continuación se resume el contenido de cada uno de los capítulos que conforman esta memoria:

- **Capítulo 2: Riesgo de Longevidad**

  En el capítulo 2 se tratarán, la historia de las tablas de mortalidad y su uso, la importancia del riesgo de longevidad en los seguros de vida, así como del riesgo de nivel tendencia y volatilidad referida a la longevidad y su tratamiento en Solvencia II.

  También se verán las principales causas de muerte en España, y la evolución de la esperanza de vida, así como las nuevas formas de protegerse de este riesgo.

- **Capítulo 3: Orígenes y funcionamiento de las Redes de Neuronas Artificiales**

  En el tercer capítulo se mostrarán las ideas generales sobre las redes de neuronas artificiales. Se mostrará la historia que está detrás de lo que hoy se conoce por Redes de Neuronas Artificiales (RNA).

  Se explicará el modelo biológico de neurona del cual parten, con el objetivo de entender mejor las ideas en las cuales se fundamentan las RNA.

  En este capítulo también se explican una serie de conceptos, en cuanto a estructura y forma de funcionamiento de una red neuronal artificial

- **Capítulo 4: Redes de neuronas de Base Radial**

  En este capítulo se explica la topología de las redes neuronales de base radial (RNBR), su funcionamiento, aplicaciones y características.

  Se detallan los algoritmos utilizados en el simulador implementado, la función de cada capa y los algoritmos que permiten que una red de este tipo sea capaz de "aprender".

- **Capítulo 5: Algoritmos genéticos**

  Capítulo dedicado al funcionamiento de los algoritmos genéticos, donde se explica su fundamentación desde el punto de vista biológico, y donde se explica la forma de funcionamiento con aquellos puntos que hay que tener en cuenta en la implementación de uno de ellos.

- **Capítulo 6: Experimentos y resultados**

  Capítulo en el que se expone la experimentación realizada, así como los resultados devueltos por la red de neuronas para cada una de las edades.

  Se comentan las conclusiones a las que se ha llegado, y unas posibles líneas de trabajos futuros.

CAPITULO 2 – RIESGO DE LONGEVIDAD

2. RIESGO DE LONGEVIDAD

2.1. UN POCO DE HISTORIA

2.2. RIESGO DE LONGEVIDAD

2.3. MODELOS DE LONGEVIDAD

2.4. SOLVENCIA II Y RIESGO DE LONGEVIDAD

2.5. MÉTODOS DE DISMINUCIÓN DEL RIESGO DE LONGEVIDAD

CAPÍTULO 2 - RIESGO DE LONGEVIDAD

En el presente capítulo se tratarán, la historia de las tablas de mortalidad y su uso, la importancia del riesgo de longevidad en los seguros de vida, así como del riesgo de nivel tendencia y volatilidad referida a la longevidad y su tratamiento en Solvencia II.

También se verán las principales causas de muerte en España, y la evolución de la esperanza de vida, así como las nuevas formas de protegerse de este riesgo.

2.1 - Un Poco de Historia

En principio, las tablas de mortalidad se pueden pensar como un instrumento de análisis demográfico, vendrían a ser modelos teóricos donde se muestra para cada grupo de edad, la probabilidad de morir antes de su próximo cumpleaños.

No sólo son utilizadas para estudios demográficos, sino también para estudios de salud pública, epidemiológicos, de madurez demográfica, ciencias actuariales, bioestadística etc.

La tabla de mortalidad de periodo, se puede entender desde el punto de vista contemporáneo, donde se estudia la mortalidad respecto a una cohorte, siendo por tanto construida de acuerdo a la población que para determinadas edades hay en un determinado momento. En este caso constituye un medio de estudio transversal de la mortalidad.

Por otro lado la tabla de mortalidad de cohorte, se puede entender desde el punto de vista de un análisis longitudinal de una generación concreta, en la cual el estudio es de cómo la mortalidad va afectando a esa generación a lo largo del tiempo. En este caso el tiempo de estudio es muy largo, pues parte desde el inicio hasta la completa extinción de aquella generación de estudio, lo hace que sean poco operativas.

Se considera que fue John Graunt (Londres, 24 de abril de 1620 - Londres, 18 de abril de 1674), estadístico inglés, y considerado el primer demógrafo, el fundador de la bioestadística y el precursor de la epidemiología, el primero que realizó sus investigaciones estadísticas, demográficas y actuariales basándose en los boletines de mortalidad (Bills of Mortality).

Por otro lado, y considerando la construcción de una tabla de mortalidad completa, se considera que fue Halley (1693) el primero en construirla bajo la hipótesis de estacionariedad de la población [Nieto y Vegas, 1993]. La estacionariedad es una característica que se asume para simplificar los cálculos, la cual implica que el tamaño de la población permanece estático y constante, suposición que ha persistido hasta la actualidad, aunque esto en la realidad raramente ocurre.

Fue con posterioridad, cuando Nicolás Titens, F. Bayly y Jorge Barret, idearon los símbolos de conmutación, los cuales permitieron agilizar los cálculos del seguro [Nieto y Vegas, 1993].

La estructura básica de una tabla de mortalidad vendría estar constituida por las siguientes columnas [Villalón, 1994]:

x: La edad del individuo, estando contenida en el intervalo $0 \leq x \leq w$
l_x: En número de individuos supervivientes a la edad x
d_x: El número de fallecidos entre la edad x y $x+1$ siendo por tanto $d_x = l_x - l_{x+1}$
q_x: Tanto anual de fallecimiento a la edad x siendo $q_x = \frac{d_x}{l_x}$
p_x: Tanto anual de supervivencia a la edad x siendo $p_x = \frac{l_{x+1}}{l_x} = 1 - q_x$

También las tablas de mortalidad suelen contener otros símbolos de conmutación, cuya misión es la de facilitar la realización de diversos cálculos actuariales, como por ejemplo el cálculo de las primas o el de las reservas. A continuación se muestran estos símbolos de conmutación:

$D_x = v^x l_x$ donde $v^x = (1+i)^{-x}$ donde $i = interés\ técnico$

$$N_x = \sum_{t=x}^{w} D_t$$

$$C_x = v^{x+1} d_x$$

$$M_x = \sum_{t=x}^{w} C_t$$

$$R_x = \sum_{t=x}^{w} M_t$$

Las tablas de mortalidad se suelen hacer para cada uno de los sexos, ya que los estudios indican claramente, que la esperanza de vida no es la misma para el sexo masculino que para el femenino.

Como se puede ver en la pagina del *INE,* la esperanza para ambos sexos al nacer, se comprueba que va aumentando año a año, lo cual se puede ver en el siguiente gráfico.

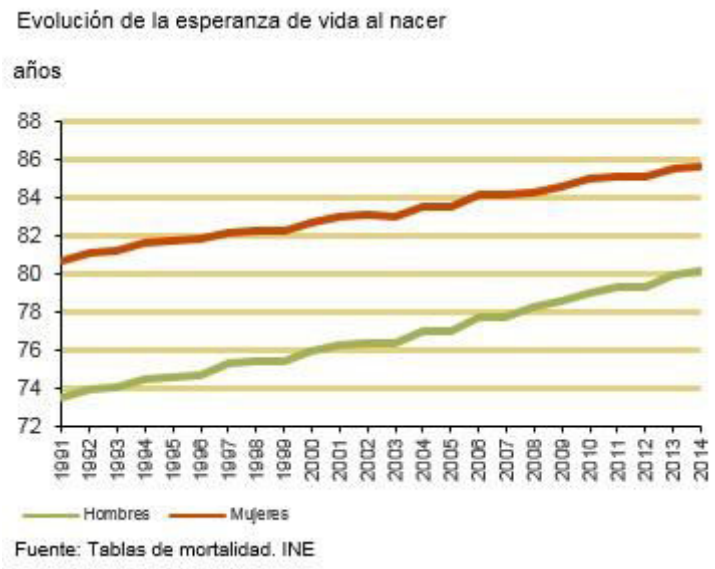


Figura 1: Esperanza de vida (Fuente INE)

En cuanto a las tablas de mortalidad, se ha visto como las tablas de periodo no tienen en cuenta este aumento en la esperanza de vida, por lo que no deberían utilizarse más allá del medio plazo. Para intentar solventar en cierta medida este problema de las tablas de mortalidad de periodo, existen otras tablas llamadas generacionales o de cohorte, en las cuales se incluye una tabla de probabilidades base, y unos factores de mejora, que

permiten ir calculando como varía esta esperanza de vida en el tiempo, por ejemplo las conocidas tablas PerM/F-2000.

2.2 - Riesgo de Longevidad

Pareciera una paradoja que vivir mucho pudiera ser un riesgo, se nos podría ocurrir que quizá desde el punto de vista de los recursos de un planeta, sí pudiera ser un problema, y seguramente si se preguntase a alguien al azar, quizá ésta sería la primera respuesta que se les ocurriría, pero ¿Quién no quiere vivir más?.

Pues no, el problema no viene referido a la cantidad de recursos disponibles, ni a problemas de superpoblación, el problema hay que verlo desde un punto de vista más crematístico.

Cuando una persona ha trabajado durante toda su vida, espera que cuando le llegue su merecido descanso, ya sea el sistema público de pensiones, o un sistema privado, le otorgue una renta de por vida, por medio de la cual pueda subsistir dignamente, hasta el final de sus días.

Si el mundo actual fuera como aquel representado en aquella película dirigida por Michael Anderson, conocida como "La fuga de Logan", no estaríamos hablando ahora del riesgo de longevidad, y por de pronto se antoja que los actuarios tendrían poco trabajo, al menos en la parte de suscripción vida.

La realidad es que la esperanza de vida de las personas en el mundo industrializado se va alargando día a día, de hecho según cálculos se alarga 15 minutos por cada hora vivida, y de hecho no se sabe si esta esperanza tocará techo en algún momento, ni mucho menos en qué momento.

Si atendemos a las palabras de *The Economist* en el artículo titulado "My Money or your life", se dice que "Riesgo de longevidad, la posibilidad de que la gente viva más tiempo de lo esperado, es potencialmente muy caro. No importa el dramático impacto de una cura para el cáncer: añadir un año extra a la esperanza de vida promedio aumenta

la factura de las pensiones del mundo un 4%, o alrededor de $ 1 billón, según el FMI." [Economist, 2014].

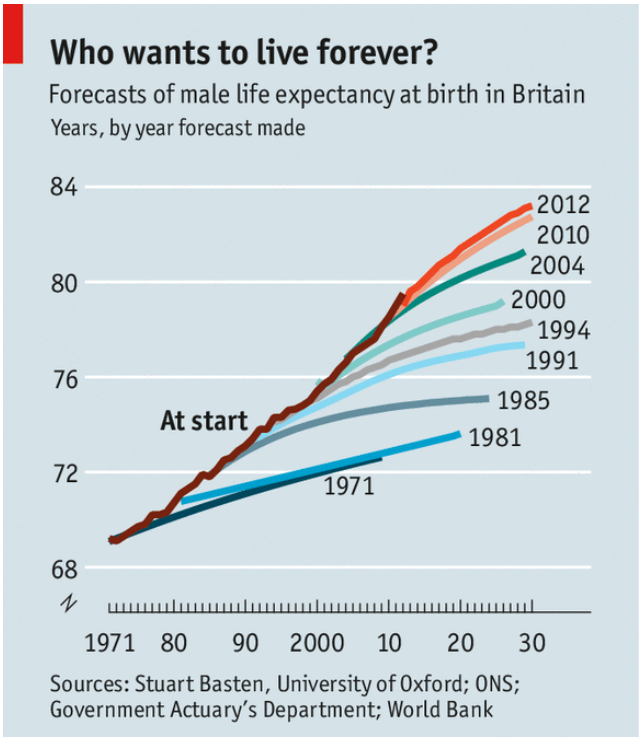


Figura 2: Fuente "The economist, agosto 2014"

El riesgo de longevidad por tanto surge del hecho de que al vivir más, los recursos para el retiro se agoten, es decir que aquellos que se han comprometido a pagarle ese retiro no puedan hacerlo por falta de recursos.

No se puede olvidar, que el pago de una renta vitalicia viene de la mano de unos cálculos basados en la matemática actuarial, la cual respaldada por una tabla de mortalidad, la edad del participe, el momento de retiro, el valor de la renta a percibir, y un interés técnico, es capaz de calcular lo que el participe deberá aportar durante su vida laboral, para percibir la renta durante el resto de la vida.

El problema viene derivado, de que la esperanza de vida que hoy se le estima a esa persona, no es la que verdaderamente tiene, ya que ésta cambia a cada momento, y no es por tanto un ente estático sino dinámico.

La persona documentada que más ha vivido hasta el momento fue la francesa Jeanne Louise Calment, que falleció a la edad de 122 años. Hay personas que aseguran haber vivido más tiempo, pero este hecho no se ha podido comprobar a ciencia cierta, como el caso de Shirali Muslimov, Azerbaiyano de supuestamente 168 años o el caso de Tuti Lusupova, de Uzbekistan de 134 años.

Este trabajo se ha realizado por medio de las tablas de mortalidad de la *Human Mortality DataBase* [HMDB] que es una web donde se pueden encontrar las tablas de mortalidad de numerosos países. En el caso de España, la longevidad máxima está fijada en los 110 años.

El caso es que no se conoce realmente las causas de que unas personas vivan más o menos, parece según los últimos estudios, que la causa de que haya personas más longevas que otras, es que han heredado el cromosoma 9p21.3, pero no se sabe si no habrá otros factores.

En biología se ha comprobado en experimentos genéticos con moscas y ratones, que alargar la vida de éstos era posible, alargando la longitud de la cadena de telómeros de sus células, permitiendo que crezca el número de divisiones posibles de una célula.

Los telómeros no son sino los extremos de los cromosomas, que marcan el número máximo de las posibles divisiones de una célula, ya que cada vez que una célula se divide los telómeros decrecen en longitud, llegando un momento que no se pueden dividir más, marcando esto el punto y final de la célula, y llegando al agotamiento vital del ser vivo.

Ahora mismo se están estudiando el papel de los telómeros en el mundo vegetal, como antes se había hecho en el mundo animal, para intentar descubrir cómo es posible que haya árboles capaces de vivir miles de años.

Por tanto, parece haber una relación directa entre la longitud de los telómeros en las células madre, con la esperanza de vida, así como con el cáncer, pues parece ser que unas cadenas cortan hacen más proclives a quienes las poseen a padecer algún tipo de cáncer en el futuro. Es paradójico no obstante, que la falta de telómeros produzca en las

células la inmortalidad, no en vano las células cancerosas, son células con un poder de división indefinido, y por ende inmortales, aunque de división descontrolada.

Las principales causas de mortalidad en España siguen siendo aquellas relacionadas con el corazón, seguidas de los accidentes cerebrovasculares, el cáncer, insuficiencias respiratorias… pero aún así la esperanza de vida sigue aumentando conforme avanza el tiempo.

Las razones de que esto ocurra no son conocidas, quizá sean las mejoras en la sanidad, pues parece que el medio ambiente no ayuda, ya que cada día está más contaminado. Por el lado de la alimentación, los efectos sobre la salud de los aditivos y transgénicos, y de sus interacciones cruzadas a lo largo del tiempo, no son bien conocidos, aunque eso sí, empieza a florecer una vertiente que aboga por llevar una alimentación más saludable, basada en alimentos ecológicos, principalmente de origen vegetal, que quizá llegue a provocar un gran salto cuantitativo en la esperanza de vida, el tiempo lo dirá.

En cualquier caso como ya se ha comentado, la esperanza de vida sigue aumentando, y el riesgo de longevidad es algo muy real que hay que tener muy en cuenta. El riesgo de longevidad se compone de otros subriesgos, como son el riesgo base o de nivel, el de tendencia y el de volatilidad.

El riesgo de nivel se produce cuando se produce un incremento o decremento de las tasas de mortalidad debido a que la media no es suficientemente explicativa. Puede ser debido a que no se cuenta con las suficientes observaciones, y pueden ayudar a solventarlas técnicas más modernas de tarificación y reservas.

En cuanto al riesgo de tendencia, éste se produce debido a la incertidumbre sobre las tasas de mortalidad a largo plazo, ya que se calcula basándose en tendencias pasadas. Se intenta solventar en cierta medida añadiendo factores de mejora en la mortalidad, teniendo en cuenta los avances médicos y como afectan a la mortalidad desde el punto de vista estadístico, en general las tendencias tardan varias décadas en ir evolucionando, por ejemplo las referentes a los estilos de alimentación, o ciertos avances relativos a la medicina.

El riesgo de volatilidad se presenta al producirse altibajos en la mortalidad de un ejercicio a otro. Esto se puede producir cuando no se cuenta con una cartera suficientemente grande, que sea capaz de diversificar el riesgo. También se puede producir cuando no hay una adecuada selección de riesgos.

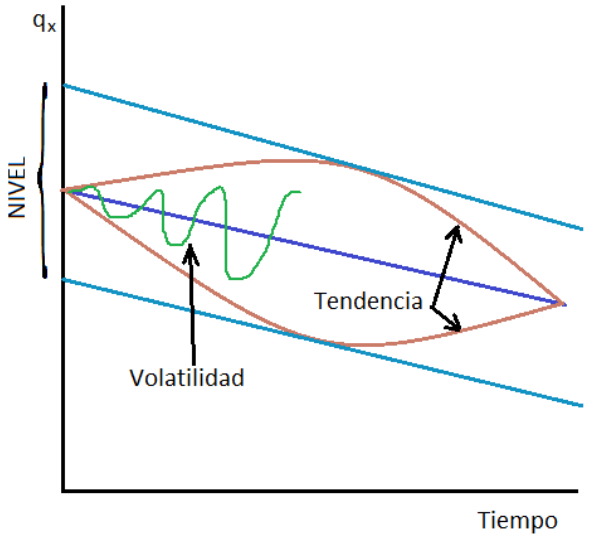


Figura 3: Riesgos de tendencia, volatilidad y nivel

2.3 - Modelos de Longevidad

No son pocos los modelos estocásticos, que pretenden modelar la mortalidad o la supervivencia del hombre, en la historia hay unos cuantos ejemplos de esto, y entre los más famosos se pueden citar unos cuantos. Los primeros que se mencionan son los más antiguos y conocidos, aumentando en nivel de complejidad según son más modernos.

La ley de Moivre supone que la función de supervivencia es lineal con la edad, donde esta ley se puede escribir como $l_x = l_0 - \frac{l_0}{\omega}x = l_0(1 - \frac{x}{\omega})$ siendo χ la edad, y ω la edad máxima a la que se puede llegar con vida. Vemos que este es un modelo muy sencillo, que presupone que las muertes son iguales todos los años, la cual coincide con la pendiente de la función de supervicencia cambiada de signo. En este caso el tanto instantáneo de mortalidad y la probabilidad de fallecimiento coinciden y son $\mu_\chi = q_\chi = \frac{1}{\omega - x}$.

La primera ley de Dormoy cuya función de supervivencia es de tipo exponencial, donde la formula es de la forma $l_\chi = K \cdot S^\chi$ donde cuando x=0 se obtiene el valor inicial de cohorte que vale K *ya que* $l_0 = K \cdot S^0 = K$, y donde para que sea una función decreciente S ha de ser un valor menor que 1. En este caso $\mu_\chi = -\ln(S)$ y $q_\chi = 1 - S$. En este caso la fuerza de mortalidad es independiente de la edad.

En la segunda ley de Dormoy, para subsanar que tanto la tasa instantánea de mortalidad, como la probabilidad anual de fallecimiento no dependan de la edad, lo cual es poco realista, se propone que $l_\chi = K S_1^\chi S_2^{\chi^2}$ donde S_1, S_2 son inferiores a 1, y al igual que en la primera ley $K = l_0$. En este caso el tanto instantáneo de mortalidad es igual a $\mu_\chi = -2\ln(S_2)\,\chi - \ln(S_1)$, el cual es creciente con la edad, pero con crecimiento relativo decreciente. Por otro lado, la probabilidad de fallecimiento anual en la segunda ley es $q_\chi = 1 - (S_1 \cdot S_2^{2\chi+1})$.

La ley de Sang es una variante de la primera ley de Dormoy, donde se añade una constante independiente de la edad quedando entonces $l_x = a + Kb^\chi$ con K positivo y $0 < b < 1$. En ese modelo el tanto instantáneo quedaría como $\mu_\chi = \frac{\ln(b)}{(b^{\omega-x}-1)}$ el cual es creciente con la edad, y por otro lado una probabilidad anual de fallecimiento de $q_x = \frac{1-b}{1-b^{\omega-\chi}}$.

La ley de Gompertz también considera un tanto instantáneo de mortalidad creciente con la edad, pero el crecimiento es un crecimiento relativo constante. En este caso la fórmula del tamaño de la población es $l_x = K g^{C^\chi} = l_0 g^{C^{\chi-1}}$ siendo $K = e^{-D} > 0$ y $g = e^{-\frac{B}{\ln C}}$ con 0<g<1 y donde B= e^h positiva, y $C = e^\gamma > 1$. Por tanto visto lo anterior el tanto instantáneo de mortalidad es $\mu_\chi = -\ln(\mathrm{g})\ln(\mathrm{C})C^\chi$ mientras que la probabilidad de fallecimiento anual quedaría como $q_\chi = 1 - g^{C^\chi(C-1)}$.

En la primera ley de Makeham, se supone que el tanto instantáneo de mortalidad se conforma de acuerdo a la expresión $\mu_\chi = A + BC^\chi$con A>0, B>0 y 0<C<1, donde el primer sumando se correspondería con las muertes de tipo accidental, mientras que el

segundo sumando se correspondería a la ley de Gompertz, lo que supone que existe una resistencia a la muerte que es decreciente con la edad, y se correspondería con las muertes por causas naturales.

En este caso $l_\chi = KS^\chi g^{C^\chi}$ $K = e^{-D} > 0$, $g = e^{-\frac{BC}{lnC}}$ cte < 1, $S = e^{-A}$, y la probabilidad de fallecimiento a un año es $q_\chi = 1 - Sg^{C^\chi(C-1)}$.

Los arriba mencionados aunque han ido aumentando en complejidad no son los más utilizados en la actualidad, aunque sí es cierto que alguno de los actuales se basan en refinamientos sobre los arriba descritos. A continuación se presentarán algunos de los modelos, que se encuentran actualmente en uso.

El modelo de Lee-Carter apareció en 1992 y es bastante usado en la proyección de la mortalidad, el tanto instantáneo de mortalidad tiene la forma $\mu_{x,t} = e^{\alpha_x + \beta_x K_t}$ siendo t el tiempo transcurrido, y x la edad. En este modelo se α_x tiene cumple el papel del efecto de la edad en la mortalidad, β_x explicaría la variación que se produce por periodo en función de la edad, y K_t describe el periodo pero con independencia de la edad.

El problema de este modelo, es que para ciertas transformaciones de los parámetros, se obtienen los mismos valores, y también que se supone que no existe dependencia entre la edad y el tiempo.

El modelo de Renshaw y Haberman (2006) generaliza el modelo anterior, añadiendo un efecto cohorte siguiendo la expresión $\mu_{x,t} = e^{\beta_x^{(1)} + \beta_x^{(2)} K_t + \beta_x^{(3)} \gamma t - x}$ y presentando también problemas con la transformación de los parámetros.

El modelo APC (Age-Period-Cohort) presentado por Currie (2006) es un modelo derivado también del modelo Lee-Carter que incluye el efecto de la edad, del periodo y de la cohorte, con $\mu_{x,t} = e^{\beta_x^{(1)} + K_t^{(2)} + \gamma_{t-x}^{(3)}}$, donde este modelo y los dos anteriores consideran la edad, el periodo y el efecto cohorte por separado, considerando aleatoriedad de un año al siguiente.

EL modelo CBD Cairns, Blake y Dowd (2006) presenta el siguiente modelo de probabilidades de fallecimiento $logit\ q(t,x) = \beta_x^{(1)} K_t^{(1)} + \beta_x^{(2)} K_t^{(2)}$, el cual no tiene el problema de identificabilidad, donde una transformación de los parámetros podía llevar a los mismos resultados. Este modelo asume una suavidad entre edades en el mismo año, pero no en diferentes años. Es usado para modelar la mortalidad a edades altas, construyéndose bajo la observación de que el logaritmo de las tasas de mortalidad son casi lineales.

Este modelo cuenta con varias variantes, como el CDB con efecto cohorte, o el CDB con efecto cohorte y componente cuadrático, pero a diferencia del anterior, éstos sí que cuentan con el problema de identificabilidad.

El problema de las edades avanzadas es que cuentan con menos exposición, ya que el tamaño de la muestra es inversamente proporcional a la edad, a edades más altas cada vez la muestra es más pequeña, y una muerte tiene mayor incidencia sobre la probabilidad de fallecimiento. Para su predicción han aparecido algunos modelos como es el de Kannistö (1992) con $\mu_x = \frac{\phi_1 e^{\phi_2 x}}{1+\phi_1 e^{\phi_2 x}}$

Por otro lado los P-Splines permiten un suavizado en dos dimensiones, en el de la edad y el periodo, y han sido utilizados por el Continuous Mortality Investigation (CMI) del Institute and Faculty of Actuaries (IFoA).

2.4 - Solvencia II y Riesgo de Longevidad

Solvencia II es el nuevo marco regulatorio que está basado en 3 pilares, siendo el pilar I el que versa sobre los requisitos de capital. En Solvencia II se calcula la necesidad de capital, basándose en el riesgo que se corre, de tal forma que se busca cubrir el 99,5% de los sucesos posibles, es decir, que la compañía pueda hacer frente a las pérdidas que se puedan producir en 199 de cada 200 veces, es decir, que se mueve en un intervalo de confianza del 99,5% o lo que es lo mismo VaR del 99,5% (Value at Risk).

Esto se puede entender como que existiría una probabilidad en 200 años de que la aseguradora se arruine, por no poder hacer frente a los pagos derivados de los siniestros producidos ese año.

En Solvencia II los activos se valoran de forma consistente al mercado, esto quiere decir que se han de valorar respecto a un mercado donde nuestro activo se compre y se venda, y que además este mercado sea líquido (muchas transacciones), y profundo (que haya muchos compradores y vendedores), el problema es que las provisiones de una aseguradora, que conforman el pasivo, y que son aquellos compromisos que debe cumplir con sus clientes, no hay un mercado líquido y profundo donde se puedan vender y comprar, por lo que se debe aproximar ese valor.

En la valoración que hay que realizar, se proyectan aquellas entradas y salidas de capital que generan los compromisos adquiridos, y se calcula la denominada mejor estimación o best estímate, y se hace la valoración actual a una tasa de descuento determinada.

Solvencia II determina que la mejor estimación no es suficiente, y que para dar mayor seguridad a los compromisos adquiridos, hay que combinarla además con un margen de riesgo, que está relacionado con el capital de solvencia SCR exigido por los compromisos afectados.

Por otro lado los fondos propios serían la diferencia entre los activos que sustentan nuestro pasivo, y nuestro pasivo, pero esos fondos propios no son libres, ya que aunque

la provisión sea un reflejo de nuestras obligaciones desde el punto de vista asegurador, este riesgo no es el único, y es posible que existan más cosas que puedan salir mal, por ejemplo que una parte de los activos se pierdan, por default de algún emisor de instrumentos financieros. Como se puede ver, hay muchos riesgos que se han de tener en cuenta en Solvencia II, para poder hacer frente a la actividad aseguradora.

El punto central en torno al que se mueve el pilar I de Solvencia II, es el cálculo del capital de solvencia, y el proceso descrito en la fórmula estándar de Solvencia II viene a resumirse en el cálculo del patrimonio libre, el cual es el patrimonio que no está sujeto a hacer frente a los compromisos del pasivo, del que debe disponerse para garantizar que se puede hacer frente a cualquier situación negativa que se pueda producir en los 12 meses siguiente, con un nivel de confianza del 99,5%.

Obviamente dentro de los riesgos que se pueden producir, hay una correlación entre -1 y 1, siendo 0 cuando no existe correlación entre ambos riesgos, y esto hay que tenerlo en cuenta, ya que es muy difícil que se vayan a producir todas las circunstancias desfavorables a la vez, con lo que el capital de solvencia sería muy alto pero irreal.

Dentro del cálculo del SCR (Solvency Capital Requeriment), se realizan una serie de cálculos basado en los tipos de riesgos que se pueden dar. Es decir que el SCR global se calcula como la suma correlacionada de otras cargas de capital basados en distintos tipos de riesgos, como son la de mercado, el de activos intangibles, el de suscripción vida, el de suscripción de salud, el de no vida, etc.

El riesgo de mercado agrupa los riesgos de tasa de interés, los de acciones, los de riesgo inmobiliario, el de cambio, el de spread o el riesgo de concentración, que como se ha comentado antes, se asume que no se van a dar todos los posibles problemas a la vez, y que los cálculos se basan en una matriz de correlaciones para este tipo de riesgo que es el de mercado.

En la figura 4, se puede ver la agrupación de estos riesgos, para llegar al cálculo del SCR global de la cartera.

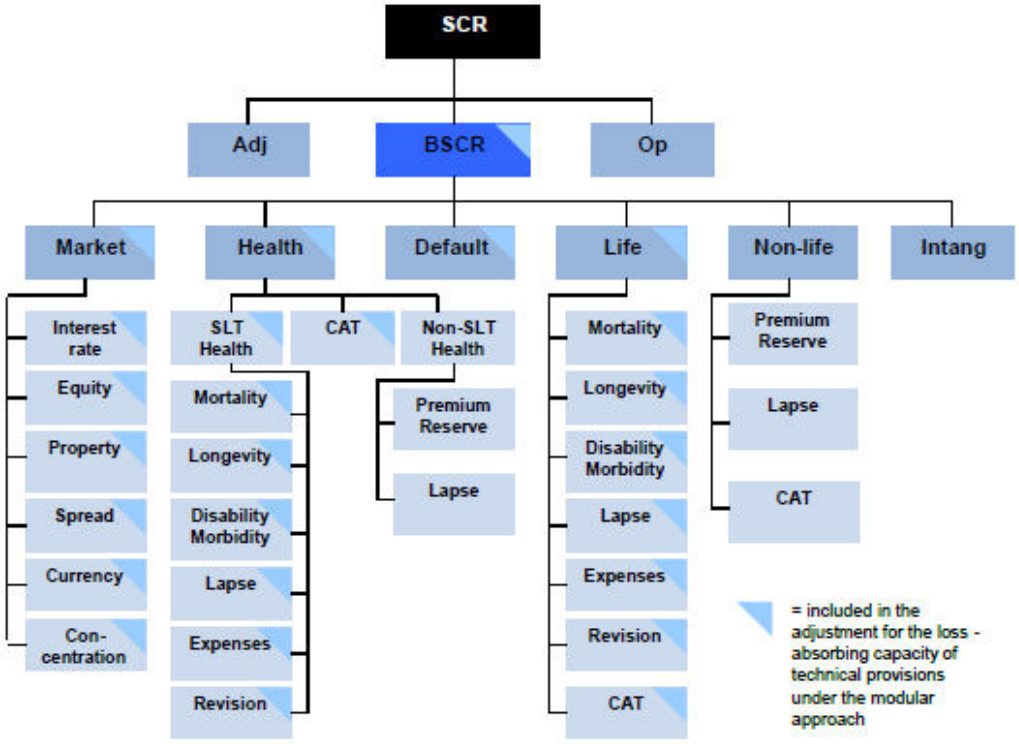


Figura 4: Gráfico con la tabla de riesgos de EIOPA (European Insurance and Occupational Pensions Authority)

En la figura anterior vemos que el riesgo de longevidad cumple un rol en el riesgo de suscripción de vida y en el de salud. El shock que Solvencia II le aplica al subriesgo de longevidad es del 20%, es decir que en cada cohorte se aplica un descenso en la mortalidad del 20%, en pocas palabras, para cada grupo de edad se producen un 20% menos de defunciones. Esto provoca obviamente que el pasivo crezca, es decir, que las obligaciones de la aseguradora sean mayores en el caso de las rentas vitalicias.

Obviamente una cartera no está conformada solo por rentas vitalicias, también hay seguros de vida tradicionales, donde el tomador se compromete a pagar la prima, y cuando fallece el asegurado, los beneficiarios obtienen una cierta cantidad previamente pactada. En este caso la longevidad es algo que beneficia a la aseguradora, pues el beneficio se encuentra en pagar cuanto más tarde mejor.

En este trabajo el estudio se centra en el riesgo de tendencia de la longevidad, y se pretende proyectar a futuro la tabla de mortalidad, pronosticada por una red de

neuronas, comprobando qué error se produce en los años de pronóstico, de acuerdo a la muestra utilizada para hacer el backtesting, y de esta forma tener una idea de hasta qué punto la proyección es más o menos confiable.

Aunque Solvencia II indica que en la fórmula estándar hay que aplicar un shock de longevidad del 20%, es cierto que se pueden utilizar modelos internos, convenientemente calculados, aunque les tiene que dar validez el regulador. Este trabajo fin de máster, intenta ofrecer otra alternativa a las ya existentes, en cuanto al pronóstico de las tablas de mortalidad, no en vano *"Del conjunto de riesgos relacionados con la vida humana que habitualmente suscribe una entidad de seguros, el más complejo de medir y gestionar es el riesgo de supervivencia"* [Rodríguez-Pardo, 2014].

2.5 - Métodos de disminución del Riesgo de Longevidad

Como se ha visto anteriormente, bajo el marco de Solvencia II, el shock de longevidad afecta al capital requerido, ya que el estrés aumenta el tamaño de la provisión, es decir de las obligaciones de la aseguradora. Según sea el volumen de este capital y/o el apetito al riesgo de la aseguradora, es posible que ésta desee que el capital requerido sea menor, o puede desear que el riesgo disminuya.

Hay que tener en cuenta también que el riesgo puede variar atendiendo de diversos factores, por ejemplo el tamaño de la cartera, a menor tamaño menos diversificada y la volatilidad será mayor, lo que incurrirá en un mayor riesgo. También habrá factores como los perfiles de riesgo de la cartera, la experiencia recogida, el riesgo normativo, etc.

La aseguradora puede tomar decisiones a la hora de mitigar el riesgo, por ejemplo si se amplía el tamaño de la cartera se introduce más aleatoriedad, habrá personas que vivan más y otras menos que compensen. Para hacer esto también hay que tener en cuenta el estilo de vida de los componentes de la cartera, ya que se ha valorado que el tipo de vida, la alimentación y el estrés pueden ser los responsables hasta del 80% de la longevidad, esto fluctúa entre unos estudios y otros, pero se ha podido comprobar por

medio del estudio de gemelos univitelinos, que el estilo de vida es más importante que los genes.

Así pues si una cartera se compone en mayor medida por personas con hábitos de vida muy saludables, contará con mayores problemas con el riesgo de longevidad, obviamente hablando desde el punto de vista de rentas vitalicias, por lo que conocer este hecho debería afectar al pricing. Por otro lado si se cuenta con una cartera de seguros de vida tradicionales, esto mitigará el riesgo de longevidad, ya que es este caso interesa que los asegurados vivan lo máximo posible, pues mortalidad y longevidad están inversamente correlacionados, y el coeficiente en Solvencia II entre ambos es del -0.25.

También hay que ser conscientes del efecto de la antiselección en la cartera, esto es que la persona que se hace un seguro de rentas vitalicias, probablemente vivirá más que la media, mientras que el que se hace un seguro de vida vivirá menos. Una opción es lanzar algún producto para edades altas, cuya esperanza de vida en principio no tendrá muchos cambios, para mitigar el riesgo de los más jóvenes.

Por otro lado, una heterogeneidad alta en las cuantías que percibirá cada uno de los beneficiarios de una cartera, tendrá un mayor riesgo al introducir una mayor volatilidad en la misma.

Otro factor importante es contar con tablas de mortalidad adecuadas, convenientemente actualizadas, habiendo países que indican la conveniencia de su actualización anual. Aquí también es importante contar con la experiencia aseguradora de la compañía, que con un histórico de pólizas conveniente, sea capaz de extraer conclusiones.

La cobertura natural también puede ser efectiva, esto no es más que en un mismo producto compensar el fallecimiento con la mortalidad.

Un producto que intenta mitigar el riesgo de longevidad o supervivencia, es el fondo de rentas vitalicias agrupadas, cuyas prestaciones están vinculadas a la experiencia de mortalidad del grupo, y de esta forma son los participantes de estos fondos los que asumen el riesgo.

Otro ejemplo de producto que intentan mitigar este riesgo, es el seguro de longevidad el cual se contrata a una edad temprana pero no se empieza a percibir como renta hasta después de la jubilación, esto es con 70 o incluso 80 años.

También existen los bonos de longevidad y mortalidad que al igual que los bonos tradicionales, es un instrumento financiero de deuda, donde los pagos de cupones dependen de la materialización de un índice de supervivencia o de mortalidad.

Otra opción son los swaps de longevidad que no es más que el reaseguro por indemnización de longevidad, no es en sí un instrumento financiero, pero sería algo análogo. Se puede hacer una transferencia total vendiendo la cartera de rentas vitalicias a otra aseguradora, o transfiriendo por ejemplo la parte del shock de longevidad donde se cubriría el pago adicional por la desviación acordada en la longevidad. En el primer caso la cedente no estaría afectada por el riesgo de contraparte, pues se hace cargo de la cartera completa desde el principio, mientras que el segundo caso sí habría riesgo de contraparte, que es que al final la reaseguradora no pague lo convenido a la cedente, y sea esta última la que tenga que pagar todo.

De esta forma, disminuyendo el riesgo de longevidad también disminuye el margen de riesgo y el capital de solvencia necesario para cumplir con Solvencia II.

CAPÍTULO 3 – ORÍGENES Y FUNCIONAMIENTO DE LAS REDES DE NEURONAS ARTIFICIALES

3. ORÍGENES Y FUNCIONAMIENTO DE LAS REDES DE NEURONAS ARTIFICIALES

3.1. UN POCO DE HISTORIA

3.2. MODELO BIOLÓGICO

3.3. REDES DE NEURONAS ARTIFICIALES

CAPÍTULO 3 - ORÍGENES Y FUNCIONAMIENTO DE LAS REDES DE NEURONAS ARTIFICIALES

En este tercer capítulo se mostrarán las ideas generales que fluyen en torno a la idea de red neuronal. Se mostrará la historia que subyace detrás de lo que hoy se conoce por *Redes de Neuronas Artificiales* (RNA). Se mostrará el modelo biológico de neurona a partir del cual nace el modelo artificial, enlazando el funcionamiento de la neurona biológica con la artificial, para un mejor entendimiento de la última.

Después se entra de lleno con las RNA definiendo una serie de conceptos, que vienen a determinar la estructura y forma de funcionamiento de una red neuronal artificial, así como algunas de las limitaciones iniciales que más tarde se solucionarían, y se explicará la forma de trabajar de las RNA.

Este trabajo fin de máster ha llevado muchas horas de programación y depuración, ya que como se comentó en las primeras páginas, se ha implementado una red de neuronas en base radial desde cero, sin utilizar ninguna librería de terceros que ya nos la diera hecha, sin más que realizar la llamada a la interfaz.

La razón de hacerse de esta manera, es porque en vista de que las pruebas probablemente fueran muchas, como así ha resultado ser, y que la flexibilidad de una caja cerrada que te dan preparada terceras personas es muy limitada, la implementación desde la base era obligada, ya que provee de la máxima flexibilidad, pues el cambio en cualquier parte del programa de la red neuronal, cuando ha sido implementada por uno mismo es posible, no así cuando ha sido implementada por otros.

El porqué se ha realizado en C/C++, sobreviene a partir de lo comentado anteriormente, y es que como la previsión inicial era que se necesitarían ingentes pruebas, se necesitaba primero, un lenguaje de programación compilado, no interpretado, pues estos últimos son varios órdenes de magnitud más lentos en general, y por otro lado que fuera un lenguaje de programación con historia, cuyos compiladores generaran lenguaje máquina optimizado para conseguir si cabe una mayor velocidad de cómputo. La aritmética de punteros se ha utilizado porque muchas veces dependiendo del compilador usado, si no

se hace así, puede ralentizar el código de tratamiento de matrices cuando se realiza por indexación matricial común, y el tratamiento de matrices es lo que primordialmente se utiliza en la implementación de una red de neuronas. Se ha desestimado introducir el código fuente como anexo, pues serían casi tantas páginas como las de este documento.

3.1 - Un poco de historia

La inteligencia artificial, en adelante IA, entendida ampliamente como el modelado y simulación de actividades cognitivas complejas (percepción, solución de problemas, razonamiento...) que caracterizan a los organismos avanzados, y en particular a los seres humanos, se separó casi desde sus inicios en dos ramas bien diferenciadas:

- Por un lado, se trató de modelar la actividad racional mediante sistemas formales de reglas, utilizando la manipulación simbólica (generalmente mediante sistemas lógicos), constituyendo quizás la rama más conocida de la IA, que se podría denominar simbólica deductiva (se postulan una serie de reglas que nos deberían permitir llegar a una solución, y el sistema resuelve el problema realizando deducciones a partir de las reglas existentes).

- Por otro lado, se desarrollaron modelos computacionales inspirados en las redes neuronales biológicas, denominados sistemas inductivos o subsimbólicos, ya que extraen la información necesaria para resolver un problema de un conjunto de ejemplos, sin necesidad de indicar las reglas que se deben seguir para llegar a la solución.

El progreso de las neurociencias, conduce a una comprensión cada vez mayor de la estructura física y lógica del cerebro, avances en la tecnología ofrecen recursos cada vez mayores que hacen posible representar estructuras cada vez más complejas, realizando cálculos a gran velocidad, muchas veces en paralelo, apoyando y fomentando así la investigación en este campo.

Se podría situar el origen de los modelos conexionistas con la definición de neurona formal dada por McCulloch y Pitts en 1943, como "*un dispositivo binario con varias entradas y salidas"*.

La historia de las RNA se podría resumir en los siguientes hechos:

- 1943: McCulloch y Pitts publicaban el artículo "*A Logical Calculus of ideas Immanent in Nervous Activity"* [MCCULLOCH & PITTS , 1943]. La colaboración de un neurobiólogo y un matemático genera un modelo abstracto de neurona, en el que la probabilidad de que una neurona se activase dependía de la señal de entrada y de la sinapsis de conexión.

- 1949: Hebb publica el libro *"The organisation of the Behavior"* [Hebb, 1949], donde introduce dos ideas fundamentales, que han influido de manera decisiva en el campo de las redes neuronales: la idea de que una percepción o concepto se representa en el cerebro por un conjunto de neuronas activas simultáneamente; y la idea de que la memoria se localiza en las conexiones entre neuronas (sinapsis).

- 1951: Marvin Minsky y Dean Edmons fabrican con dispositivos mecánicos una máquina capaz de aprender. Para ello se basaron en las ideas de McCulloch y Pitts.

- 1956: Organizada por Minsky, John McCarthy, Nathaniel Rochester y Claude Shannon se celebró la primera conferencia sobre inteligencia artificial.

- 1958: Frank Rosenblatt desarrolla el perceptron. Un sistema que permitía interpretar patrones tanto abstractos como geométricos. El perceptron supone un gran avance en IA y el inicio de las RNA.

- 1959: Widrow y Hoff desarrollan un algoritmo muy conocido que es el denominado *regla delta* o regla del mínimo error cuadrado (LMS Error: *Least-Mean-Squared Error),*que permite cuantificar el error global cometido, y que

se conoce también como regla de *Widrow-Hoff* siendo utilizada en el modelo llamado ADALINE (Adaptative Linear Element) con una única neurona de salida, y el MADALINE (Multiple Adaline) el cual puede tener varias neuronas de salida. Este modelo de RNA es capaz de clasificar los datos en espacios separables linealmente. La red ADALINE era extremadamente similar al perceptron y pronto se generalizan estos resultados al modelo de Ronsenblatt ya que los resultados eran extrapolables. Esta red se usó por primera vez en un problema importante del mundo real, como fue la creación de un filtro adaptativo para la cancelación de ecos.

- 1969: Minsky y Papert publican el libro llamado *"Perceptrons"* [Minsky & Papert, 1969] en el que presentan el principal problema del perceptron, el famoso problema de la separabilidad no lineal ilustrado mediante la función XOR u OR Exclusivo. Este libro desmoralizó a todos los investigadores y provocó un descenso en el interés en las RNA. En la época posterior al libro de Minsky y Papert el interés generalizado por las RNA había desaparecido, sin embargo algunos investigadores decidieron continuar con las investigaciones. En este periodo aparecen los modelos de Anderson, Kohonen, Grossberg, Hopfield, etc.

- 1977: Anderson estudia y desarrolla modelos de memorias asociativas. Destaca el autoasociador lineal conocido como modelo brain-state-in-a-box (BSB).

- 1982: Hopfield elabora un modelo de RNA consistente en unidades de proceso interconectadas que alcanzan mínimos energéticos, aplicando los principios de estabilidad desarrollados por Grossberg. El modelo de Hopfield resultó muy ilustrativo sobre los mecanismos de almacenamiento y recuperación de la memoria. Su entusiasmo y claridad de presentación dieron un nuevo impulso al campo de las RNA y provocaron el incremento de las investigaciones.

- 1984: Kohonen continua el trabajo de Anderson y desarrolla modelos de aprendizaje competitivo basados en el principio de inhibición lateral. Su principal aportación consiste en un procedimiento para conseguir que unidades

físicamente adyacentes aprendieran a representar patrones de entrada similares; a las redes basadas en este procedimiento se las llama redes de Kohonen [Kohonen, 2001].

- 1986: Rumelhart, McClelland y el PDP (Parallel Distributed Processing) grupo fundado por los anteriores investigadores, y dedicado al estudio del conocimiento, editaron el libro *"Parallel Distributed Processing: Exploration in the Microstructures of Cognition"* [Rumelhart & McClelland, 1986]. Este libro supuso una revolución dentro de las RNA; en él se exponía el método Back-Propagation que resolvía el problema de la separablidad no lineal expuesto por Minsky. Después de este libro se produjo una explosión en las RNA apareciendo modelos, técnicas, campos de aplicación y fusiones híbridas de modelos. Esta explosión e interés en el tema ha durado hasta la actualidad donde las RNA son una de las herramientas más utilizadas.

- 1987: Grossberg realizó un importante trabajo teórico-matemático tratando de basarse en principios fisiológicos; aportó importantes innovaciones con su modelo ART (Adaptative Resonance Theory) [Grossberg, 1987], y junto a Cohen elabora un importante teorema sobre la estabilidad de las redes recurrentes en términos de una función de energía.

3.2 - Modelo Biológico

La teoría y el modelado de las RNA, está inspirada en la estructura y funcionamiento del sistema nervioso, donde la neurona es el elemento fundamental. Este funcionamiento artificial no deja de ser un modelo de funcionamiento basado en el funcionamiento biológico, y hoy en día se desconocen muchos aspectos del funcionamiento real de las neuronas, por lo que una red de neuronas artificiales, no deja de ser una simplificación del funcionamiento de un cerebro real.

Lo importante de los modelos artificiales es que presentan comportamientos útiles, de forma que son capaces de "aprender", reconocer y aplicar relaciones entre objetos

propios del mundo real. En este sentido, ofrecen un nuevo conjunto de herramientas que podrán utilizarse para resolver distintos tipos de problemas, y los más optimistas piensan que su estudio, probablemente nos ayudará a comprender mejor el funcionamiento de los modelos biológicos.

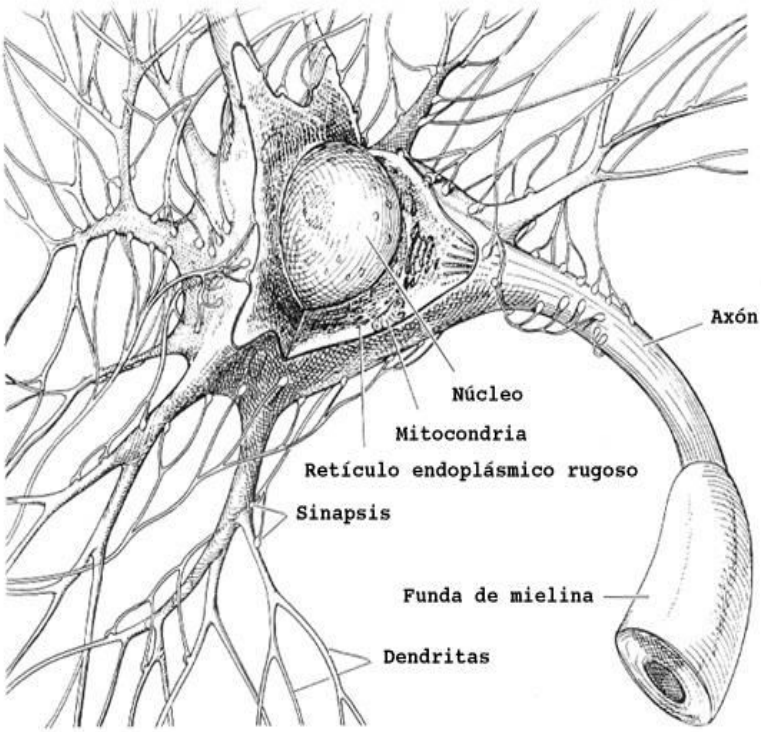


Figura 5: Partes de una neurona biológica

Una neurona es una célula viva, y como tal contiene los mismos elementos que forman parte de todas las células biológicas. Además contiene una serie de elementos distintivos y que le son característicos, como son un cuerpo celular más o menos esférico, con un tamaño de 5 a 10 micras de diámetro, del que salen una rama principal, el axón, y varias ramas más cortas, llamadas dendritas. A su vez, el axón puede producir ramas en torno a su punto de arranque, y con frecuencia se ramifica extensamente cerca de su extremo.

La característica principal que diferencian a las neuronas del resto de las células vivas, es su capacidad de comunicarse. Este proceso lo realizan en términos generales de la siguiente manera: las dendritas y el cuerpo celular reciben señales de entrada; el cuerpo celular las combina e integra y emite señales de salida. El axón transporta esas señales a

los terminales axónicos, que se encargan de distribuir información a un nuevo conjunto de neuronas. Por lo general, una neurona recibe información de miles de otras neuronas y, a su vez, envía información a miles de neuronas más. Se calcula que en el cerebro humano existen del orden de 10^{11} neuronas con 10^{15} conexiones.

3.2.1 - Naturaleza bioeléctrica de la neurona

Las señales que utilizan las neuronas son de naturaleza eléctrica y química. La señal generada por la neurona y transportada a lo largo del axón es un impulso eléctrico, mientras que la señal que se transmite entre los terminales axónicos de una neurona y las dendritas de las neuronas siguientes es de origen químico; concretamente, se realiza mediante moléculas de sustancias transmisoras (neurotransmisores) que fluyen a través de unos contactos especiales, llamados sinapsis, que tienen la función de receptor y están localizados entre los terminales axónicos y las dendritas de la neurona siguiente (espacio sináptico, entre 50 y 200 Angstroms).

3.2.2 - Señal eléctrica

La membrana de la neurona, separa el plasma intracelular del fluido intersticial que se encuentra fuera de la célula. La membrana es permeable para ciertas especies iónicas, y actúa de tal forma que se mantenga una diferencia de potencial entre el fluido intracelular y el fluido extracelular. La diferencia más notable, se da en relación con la concentración de los iones sodio y potasio. El medio externo es unas 10 veces más rico en sodio que el interno, mientras que el medio interno es unas 10 veces más rico en potasio que el externo. Esta diferencia de concentración en iones sodio y potasio a cada lado de la membrana, produce una diferencia de potencial de aproximadamente 70 milivoltios negativa, en el interior de la célula. Es lo que se llama potencial de reposo de la célula nerviosa.

La llegada de señales procedentes de otras neuronas a través de las dendritas, (recepción de neurotransmisores) actúa acumulativamente, bajando ligeramente el valor del potencial de reposo. Dicho potencial, modifica la permeabilidad de la membrana, de

manera que cuando llega a cierto valor crítico comienza una entrada masiva de iones sodio, que invierten la polaridad de la membrana.

La inversión del voltaje de la cara interior de la membrana, cierra el paso a los iones sodio y abre el paso a los iones potasio, hasta que se restablece el equilibrio en reposo. La inversión del voltaje, conocida como potencial de acción, se propaga a lo largo del axón y, a su vez, provoca la emisión de los neurotransmisores en los terminales axónicos.

Después de un pequeño periodo refractario, puede seguir un segundo impulso. El resultado de todo esto, es la emisión por parte de la neurona, de trenes de impulsos cuya frecuencia varía en función (entre otros factores) de la cantidad de neurotransmisores recibidos.

3.2.3 - Señal química

La acción química que se produce en los receptores, da lugar a cambios de permeabilidad de la membrana postsináptica para ciertas especias iónicas. Existen dos tipos de efectos locales en la sinapsis:

- El efecto excitador, cuyos neurotransmisores provocan disminuciones de potencial en la membrana de la célula postsináptica, facilitando la generación de impulsos a mayor velocidad.

- El efecto inhibidor, cuyos neurotransmisores tienden a estabilizar el potencial de la membrana, dificultando la emisión de impulsos.

Estos dos efectos, actúan sólo a lo largo de una pequeña distancia hacia el interior de la célula; se suman en el montículo del axón; la suma de los efectos excitadores e inhibidores determina si la célula será o no estimulada es decir, si emitirá o no un tren de impulsos y a qué velocidad.

3.3 - Redes Neuronales Artificiales

Las RNA son una simulación abstracta de los sistemas nerviosos biológicos, formados por un conjunto de unidades llamadas "neuronas" o "nodos" conectadas unas con otras. Estas conexiones, tienen una gran semejanza con los axones y dendritas en los sistemas nerviosos biológicos.

En general las RNA tienen una serie de características, que las hacen ser muy utilizadas frente a otros métodos, siendo estas características las siguientes:

- **Aprendizaje adaptativo:** Es la capacidad de aprender a realizar una tarea, basada en un entrenamiento o una experiencia inicial. Las RNA's tienen la capacidad de aprender continuamente, e ir adaptandose a la nueva información.

- **Autoorganización:** Una red neuronal, puede crear su propia organización o representación de la información, que recibe en la etapa de aprendizaje.

- **Tolerancia a fallos:** La destrucción parcial de una RNA conduce a una degradación parcial en su funcionamiento; sin embargo algunas capacidades de la red se pueden retener incluso sufriendo daños mayores.

- **Operación en tiempo real:** El cómputo neuronal puede realizarse en paralelo. Se utilizan ciertas tecnologías para la implementación de RNA's, de forma que se facilite este tipo de funcionamiento paralelo. La paralelización de una red neuronal permite su funcionamiento en tiempo real.

- **Facilidad de simulación y construcción:** Las RNA's son estructuras muy sencillas que permiten su fácil implementación y simulación en cualquier computador.

- **Generalización:** Si se elige bien la estructura y se produce un entrenamiento adecuado, estas redes son capaces de generalizar.

3.3.1 - Definiciones

A continuación se presentan algunas de las definiciones más conocidas de red de neuronas artificial:

- *"Colecciones de procesadores paralelos conectados entre sí en forma de grafo dirigido"* [Freeman & Skapura, 93].

- *"Las redes neuronales son modelos del proceso cognoscitivo del cerebro"* [Blum, 92].

- *"Un sistema de computación hecho por un gran número de elementos simples, elementos de proceso muy interconectados, los cuales procesan información por medio de su estado dinámico como respuesta a entradas externas."* [Hecht-Nielsen, 88].

- *"Una red neuronal es un procesador masivamente paralelo distribuido que es propenso por naturaleza a almacenar conocimiento experimental y hacerlo disponible para su uso"* [Haykin, 94]. Este mecanismo se parece al cerebro en dos aspectos:

    - El conocimiento es adquirido por la red a través de un proceso que se denomina aprendizaje.

    - El conocimiento se almacena mediante la modificación de la fuerza o peso sináptico de las distintas uniones entre neuronas.

- *"Una red neuronal es un modelo computacional con un conjunto de propiedades específicas, como son la habilidad de adaptarse o aprender, generalizar u organizar la información, todo ello basado en un procesamiento eminentemente paralelo"* [Kröse & Smagt, 93].

Se puede observar que en las definiciones mostradas más arriba se hace énfasis en que las RNA son elementos de proceso muy interconectados; no se debe olvidar que el cerebro, por ejemplo el cerebro humano, consta de aproximadamente 10^{11} neuronas, y existen aproximadamente 10^{15} conexiones, lo cual quiere decir que aproximadamente cada neurona está conectada a 10^4 neuronas trabajando todas en paralelo.

Otro de los puntos clave en las definiciones, es que una RNA es capaz de aprender, estando el conocimiento en los pesos sinápticos lo cual se explicará en posteriores secciones. Esto no es más que lo que indicó Hebb cuando dijo que la memoria se localizaba en las sinapsis de las neuronas biológicas.

Básicamente las redes neuronales artificiales, son un conjunto de unidades de proceso muy interconectadas, que trabajan en paralelo y que a partir de la interacción con unos datos de entrada en un proceso denominado *de aprendizaje* son capaces de proporcionar salidas "razonables".

3.3.2 - Estructura de una RNA

Las RNA se modelan mediante unidades de proceso: las *neuronas*. Generalmente, en cada red neuronal se pueden encontrar tres tipos de neuronas, atendiendo al modelo biológico que se estructura en las siguientes capas:

- **Capa de entrada:** Las neuronas que reciben estímulos del exterior, que serían aquellas que de alguna forma estarían relacionadas con el aparato sensorial, que toman unos datos de entrada para ser procesados.

- **Capas ocultas:** La información de las neuronas anteriores, se transmite a otras neuronas que se ocupan de su procesado. En este nivel es donde se produce el aprendizaje, por medio de una representación interna de los datos en las neuronas y sinapsis.

- **Capa de salida:** Una vez que se ha procesado toda la información, esta información pasa a las unidades de salida que son las que dan la salida del sistema.

Cada unidad de proceso o *neurona* se compone de una red de conexiones de entrada, una función de red (de propagación) encargada de computar la entrada total combinada de todas las conexiones que le llegan, un núcleo central de proceso encargado de aplicar una función denominada función de activación, y por último una función de salida que normalmente no realiza ningún proceso, con lo que la salida sería el valor devuelto por la función de activación.

En la figura 6 se puede ver un ejemplo de neurona artificial, con todos los elementos explicados anteriormente.

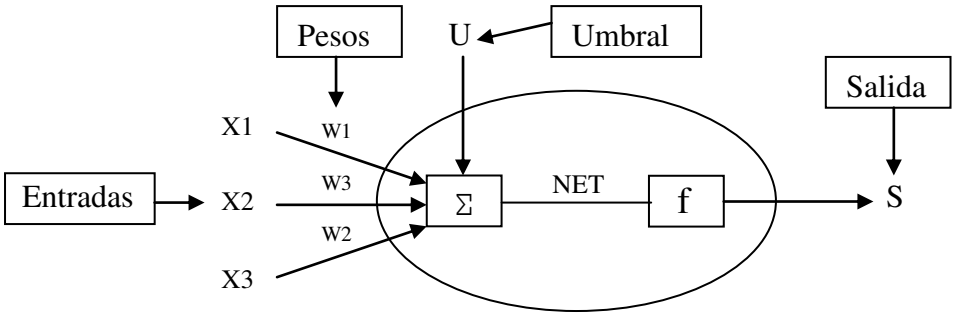


Figura 6: Neurona artificial

Redes Neuronales Biológicas	*Redes Neuronales Artificiales*
Neuronas	Unidades de proceso
Conexiones sinápticas	Conexiones ponderadas
Efectividad de las sinápsis	Peso de las conexiones
Efecto excitatorio o inhibitorio de una conexión	Signo del peso de una conexión
Efecto combinado de las sinapsis	Función de propagación o de red
Activación -> tasa de disparo	Función de activación -> Salida

Tabla 1*:* Comparación entre las neuronas biológicas reales y las unidades de proceso artificiales

En la tabla 1 se muestran los elementos que se corresponden con el modelo artificial de red neuronal, y sus homólogas biológicas.

3.3.3 - Función de red o propagación

La función de red (de propagación ó de base), es normalmente el sumatorio de todas las entradas que le llegan multiplicadas por sus pesos respectivos, es decir:

$$Net_j = \sum_{i=1}^{n} x_i w_{ji}$$

donde *j* es la unidad de proceso *j*, w_{ij} es el peso de la conexión que une la neurona i con la j de una capa posterior, y x_i es la salida de la neurona i, que es de una capa anterior. Cuando esta función tiene la forma anterior, se la suele denominar *función de base lineal.*

La función de red también puede ser otro tipo de función, como por ejemplo la función de base radial, la cual es una función de tipo hiperesférico, de segundo orden, no lineal. El valor de red, representa la distancia de la neurona a un determinado patrón de referencia, es decir:

$$Net_j = \sqrt{\sum_{i=1}^{n} (x_i - w_{ij})^2}$$

Esta función de red o de propagación, si se hace referencia al modelo biológico correspondería con la combinación de las señales excitatorias o inhibitorias que llegan a la neurona.

3.3.4 - Función de activación

La función de activación, es quizás la característica principal o definitoria de las neuronas, la que mejor define el comportamiento de la misma. Se usan diferentes tipos de funciones, desde simples funciones de umbral a funciones no lineales.

La función de activación, se encarga de calcular el nivel o estado de activación de la neurona en función de la entrada total:

$$y_i(t) = F(NET_i)$$

Entre las funciones de activación, se suele distinguir entre funciones lineales, en las que la salida es proporcional a la entrada; funciones de umbral, en las cuales la salida es un valor discreto (típicamente binario 0/1) que depende de si la estimulación total supera o no un determinado valor de umbral; y funciones no lineales, no proporcionales a la entrada (ver figura 7).

Casi todos los avances recientes en conexionismo, se atribuyen a arquitecturas multicapa, que utilizan funciones de activación no lineales como una función de umbral, una gaussiana ó en la mayoría de los casos una función sigmoidal [Quinlan, 91]. El problema de trabajar con modelos no lineales radica en que son difíciles de describir en términos lógicos o matemáticos convencionales [Rumelhart & McClelland, 86].

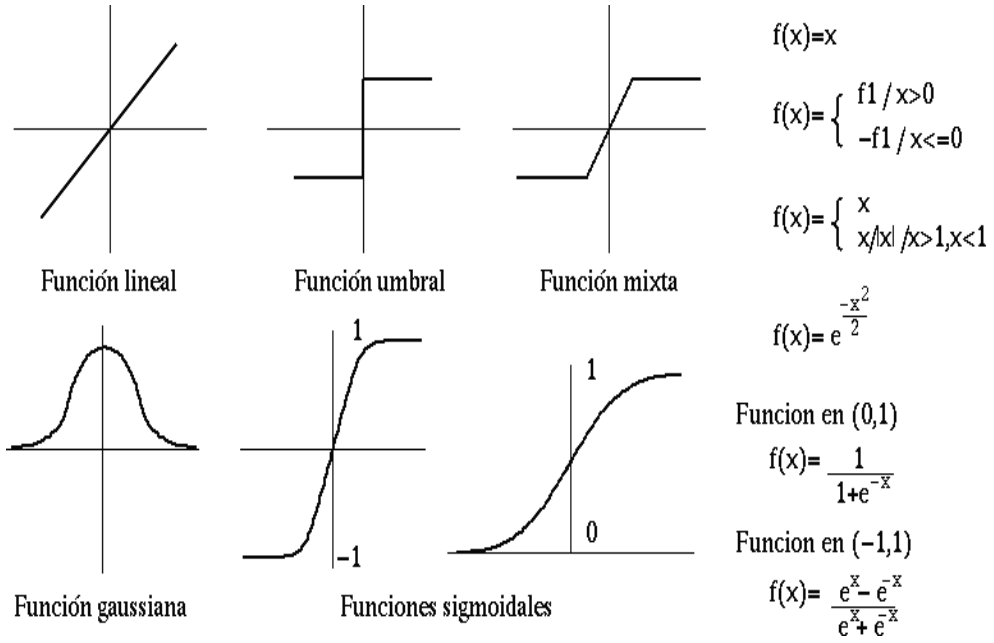


Figura 7: Tipos de funciones de activación

3.3.4.1 - Función escalón o umbral

En un principio, se pensó que las neuronas usaban una función de umbral, es decir, que permanecían inactivas y se activaban sólo si la estimulación total superaba cierto valor límite; esto se puede modelar con una función escalón: la más típica es el escalón unitario: la función devuelve 0 por debajo del valor crítico (umbral) y 1 por encima.

Después se comprobó que las neuronas emitían impulsos de actividad eléctrica con una frecuencia variable, dependiendo de la intensidad de la estimulación recibida, y que tenían cierta actividad hasta en reposo, con estimulación nula. Estos descubrimientos llevaron al uso de funciones no lineales con esas características, como la función sigmoidal, con un perfil parecido al escalón de una función de umbral pero continua.

3.3.4.2 - Función sigmoidal

La función sigmoidal es parecida a la función escalón o umbral, en cuanto a que la pendiente es elevada, y por tanto la mayoría los valores devueltos por la función, estarán comprendidos entre la zona alta o baja del sigmoide.

La importancia de esta función u otra similar, es que su derivada es siempre positiva, pudiéndose utilizar las reglas de aprendizaje definidas para las funciones de tipo escalón, con la ventaja de que la derivada está definida en todo el intervalo, lo que ayuda a la hora de utilizar los métodos de aprendizaje que utilizan derivadas.

3.3.4.3 - Función gaussiana

En este tipo de funciones se pueden adaptar tanto los centros como la anchura de la campana, lo que hace de este tipo de funciones que sean más adaptativas que las funciones sigmoidales.

Mientras que con funciones sigmoidales se suele requerir el uso de varias capas de neuronas ocultas, con el uso de este tipo de funciones suele bastar con una sola capa .

3.3.4.4 - Función mixta

Es una mezcla entre la función escalón y la función lineal, de forma que en cierto intervalo la función es lineal mientras que cuando alcanza un límite superior la activación se define como 1, y cuando sobrepasa el límite inferior la activación es 0 ó -1.

2.3.5 - Sinapsis entre neuronas

Son conexiones ponderadas que hacen el papel de las conexiones sinápticas, el peso de la conexión equivale a la fuerza o efectividad de la sinapsis. La existencia de conexiones determina si es posible que una unidad influya sobre otra. El valor de los pesos y el signo de los mismos, definen el tipo (excitatorio/inhibitorio) y la intensidad de la influencia.

2.3.6 - Formas de conexión de las redes neuronales

Las neuronas de una RNA se conectan entre sí, en lo que serían las sinapsis de sus homólogas biológicas, de tal forma que se pueden dar varios tipos de interconexiones, por ejemplo las conexiones en las que la salida de una neurona es entrada de sí misma, se dice que son *conexiones autorrecurrentes.*

Cuando se tiene una red, donde ninguna de las salidas de las neuronas que la componen, es entrada de alguna neurona de niveles anteriores, se dice que se trata de una red de *propagación hacia adelante o feedfordward*.

Cuando las salidas de las neuronas, se conectan con neuronas de niveles anteriores o del mismo nivel (incluyéndose ellas mismas), se dice que se trata de una red de *propagación hacia atrás o feedback.*

Si la salida de una neurona es entrada de otra neurona del mismo nivel se habla de *conexiones laterales*.

En una red neuronal pueden coexistir varios tipos de conexiones, no tienen porqué ser todas del mismo tipo.

2.3.7 - Mecanismo de aprendizaje

Las RNA son capaces de aprender de acuerdo a unos mecanismos, llamados mecanismos de aprendizaje. Este mecanismo o proceso, se basa en la modificación de los pesos de las neuronas, como respuesta a la información que se le suministra como entrada.

En los sistemas biológicos, se ha demostrado que las sinapsis entre neuronas se crean y se destruyen de acuerdo a su uso. Por ejemplo se ha podido comprobar en el caso de músicos, que el área cerebral encargada del ritmo, así como de la coordinación motora de los dedos, adquiere una mayor densidad con los años de práctica, ya que es necesaria una mayor coordinación que en el caso de no tocar ningún instrumento.

En las RNA la creación de una nueva conexión, implica que el peso ha de ser distinto de 0, mientras que la destrucción de una determinada conexión implica que el peso pasa a ser 0.

En el proceso de aprendizaje, los pesos de las conexiones van variando hasta que la red termina de aprender, lo cual se produce cuando los pesos son estables, es decir cuando $\partial w_{ij} / \partial t = 0$ es decir cuando la derivada de los pesos respecto al tiempo pasa a ser 0, o lo que es lo mismo, cuando el valor de un peso en el tiempo t y en $t+1$ es el mismo.

Es importante conocer el mecanismo según el cual se modifican los pesos de la red, conociéndose este mecanismo como *regla de aprendizaje.* Las reglas de aprendizaje se engloban normalmente en dos tipos de reglas, las reglas de *tipo supervisado,* y las de *tipo no supervisado.* La diferencia entre ambos tipos de aprendizaje se basa en la existencia o no de un supervisor, que es un agente externo que controla el proceso de aprendizaje de la red.

Otro criterio utilizado es el de si la red tiene una fase de aprendizaje determinada o por el contrario aprende durante todo el tiempo. Si la red tiene una fase de aprendizaje en la cual no hace nada más que aprender, y terminada ésta no retoma el aprendizaje, entonces se trata de una red de aprendizaje *OFF LINE.* Si por el contrario la red está aprendiendo de forma continua durante todo su funcionamiento, entonces se habla de redes de aprendizaje *ON LINE.*

En algunos casos se dan ambos tipos de aprendizaje, coexistiendo el aprendizaje de tipo *OFF LINE*, y el aprendizaje de tipo *ON LINE.*

3.3.7.1 - Redes con aprendizaje supervisado

Este tipo de redes, se distingue porque cuenta con un agente que dirige la fase de entrenamiento, indicando en cada momento la respuesta que debería dar la red en base a las entradas recibidas por la red.

De acuerdo a la salida que devuelva la red, el agente la comparará con la salida deseada, y procederá a cambiar los pesos de las conexiones de la red, de tal forma que la salida de la red y la deseada sean lo más parecidas posible.

Dentro de este tipo de aprendizaje se pueden distinguir además tres tipos de llevarlo a cabo que son los siguientes:

- Aprendizaje por corrección de error.
- Aprendizaje por refuerzo.
- Aprendizaje estocástico.

Aprendizaje por corrección del error

El aprendizaje por corrección del error se basa en ajustar los pesos de las conexiones de la red, en función de la diferencia entre los valores deseados y los obtenidos en la salida de la red, es decir, en función del error cometido. Entre estos algoritmos se encuentran el utilizado en la red *Perceptron* [Rosenblatt, 58], la *regla delta* o regla del mínimo error cuadrado [Widrow, 60], utilizado en las redes *Adaline y Madaline,* y que supera al anterior algoritmo en rapidez de aprendizaje.

Otro algoritmo es el de la *regla delta generalizada* conocida también por *algoritmo de retropropagación del error* (error backpropagation) [Werbos, 74], que es utilizado en redes multicapa de tipo *feedforward.* Este último algoritmo amplía el campo de aplicación de las RNA pero como desventaja se puede decir que es un algoritmo mucho más lento.

Aprendizaje por refuerzo

El aprendizaje por refuerzo, se basa en no indicar exactamente la salida que se desea, es decir que sólo se indica por medio de una señal (éxito = +1, fracaso = -1) si la salida obtenida en la red se ajusta a la deseada, y por medio de eso se ajustan los pesos de la red mediante un mecanismo basado en probabilidades. Este tipo de aprendizaje es más lento que el anterior. Como ejemplo de este algoritmo se puede citar el *Linear-Reward-Penalty* (algoritmo lineal con recompensa y penalización) [Narendra & Thathchar, 74], o el también conocido como *Adaptive Heuristic Critic* [Barto, Sutton & Anderson, 83].

Aprendizaje estocástico

Por último, el aprendizaje estocástico consiste fundamentalmente, en la realización de cambios aleatorios en los valores de los pesos de las conexiones de la red, y evaluar su efecto a partir del objetivo deseado, y de distribuciones de probabilidad. Una vez que se han producido los cambios, éstos se evalúan y si el comportamiento de la red es mejor se acepta el cambio, y si es peor se acepta en función de una determinada y preestablecida distribución de probabilidades.

3.3.9.2 - Redes con aprendizaje no supervisado

Las redes con aprendizaje no supervisado, no requieren del uso de ningún agente externo que regule el ajuste de los pesos de las conexiones entre sus neuronas. Al no recibir información sobre la veracidad de sus salidas, se dice que este tipo de redes son capaces de autoorganizarse.

La interpretación de las salidas de este tipo de redes depende de su estructura y del algoritmo de aprendizaje utilizado.

En unos casos, la salida representa el grado de similitud entre la información mostrada a la entrada en el momento actual, y la que se le había presentado hasta entonces.

En otros casos, se podría considerar un proceso de establecimiento de categorías (clustering) donde la salida sería la categoría a la que pertenece la información suministrada a la entrada.

Entre los algoritmos de aprendizaje no supervisado se suelen considerar generalmente los dos siguientes:

- Aprendizaje hebbiano.
- Aprendizaje competitivo y cooperativo.

Aprendizaje hebbiano

El aprendizaje hebbiano consiste en el ajuste de los pesos de las conexiones, de acuerdo con la correlación de los valores de activación, es decir que cuando dos neuronas conectadas dan valores positivos, se refuerza la conexión, mientras que si una da un valor positivo y la otra uno negativo, en ese caso la conexión se vería debilitada. Este tipo de funcionamiento se basa en un postulado formulado por Donald O. Hebb [Hebb, 49] en el que se indicaba que cuando una neurona tomaba parte de forma persistente en la activación de otra neurona, entonces la conexión entre ambas se veía reforzada mediante algún tipo de cambio metabólico.

Aprendizaje competitivo

En el aprendizaje competitivo, se dice que las neuronas compiten y cooperan a fin de llevar a cabo una cierta tarea. En este aprendizaje se pretende que al presentarse cierta información a la red, sólo una de las neuronas de salida se active (alcance el valor de respuesta máximo), o una por cierto grupo de neuronas.

La competición en este tipo de redes, se produce en todas las capas de la red, no sólo en la de salida, y las conexiones de este tipo de redes son recurrentes de autoexcitación, y

conexiones de inhibición por parte de las neuronas vecinas. Si el aprendizaje es cooperativo, las conexiones vecinas serán de excitación.

Con este tipo de aprendizaje, se pretende categorizar los datos que se introducen en la red, con lo que informaciones parecidas darán lugar a que se active la misma neurona de salida.

3.3.7.3 - Redes con aprendizaje híbrido

Este tipo de redes utilizan ambos tipos de aprendizaje, utilizándose cada uno de ellos en las distintas capas que conforman la red, o utilizándose en las distintas fases del aprendizaje.

Por ejemplo, el simulador implementado dispone de dos tipos de aprendizaje, uno es el aprendizaje supervisado y otro es el aprendizaje híbrido. Mientras que en el aprendizaje híbrido se determinan centros y desviaciones antes de la fase de entrenamiento en sí, en la que se modifican los pesos, en el aprendizaje supervisado tanto centros, desviaciones y pesos son modificados en la propia fase de aprendizaje de acuerdo a ciertos algoritmos y claro está, dependiendo de la información de entrada a la red.

Se ha programado un tercer tipo de aprendizaje, que es una mezcla del híbrido y del totalmente supervisado, en el que se actúa primeramente como en el caso del híbrido, calculando centros y desviaciones, pero luego en la fase de entrenamiento se modifican tanto pesos, como desviaciones, como centros, en este caso lo que se pretendía era dar una inicialización a centros y desviaciones, más allá de una inicialización aleatoria.

CAPITULO 4 – REDES DE NEURONAS DE BASE RADIAL

4. REDES NEURONALES DE BASE RADIAL

4.1. ESTRUCTURA DE LAS RNBR

4.2. ENTRENAMIENTO DE LAS RNBR

4.3. APRENDIZAJE HÍBRIDO

4.4. APRENDIZAJE SUPERVISADO

4.5. ALGUNAS NOCIONES SOBRE EL APRENDIZAJE

4.6. PRINCIPALES CARACTERÍSTICAS DE LAS RNBR´s

CAPÍTULO 4 - REDES DE NEURONAS DE BASE RADIAL

En este capítulo, se explicará la topología de las redes neuronales de base radial (RNBR), así como su forma de funcionamiento, aplicaciones y características que la hacen distintivas de otro tipo de redes como el perceptron multicapa.

Se verán también los algoritmos utilizados en el simulador implementado, tanto los referentes al aprendizaje supervisado como al método híbrido.

4.1 - Estructura de las RNBR

Las RNBR se componen de tres capas (figura 8), la primera capa o capa de entrada no realiza ningún procesado de los datos de entrada, pasándolos tal cual a la siguiente capa; es por esta razón por la que algunos autores consideran a estas redes como de dos capas, sólo considerando aquellas capas que realizan algún tipo de procesado de los datos. La capa de entrada simplemente realiza una función de simple transmisión de información.

La siguiente capa es la capa oculta a la que le llegan los datos de la capa de entrada, de tal forma que aplicando alguna función de base radial obtiene una salida que pasa a la siguiente capa. La función más utilizada es la gaussiana aunque cómo se verá existen otras funciones que se pueden utilizar. La activación de las neuronas de esta capa es de carácter local, aunque este hecho depende el tipo de aprendizaje utilizado, pero esto será explicado más adelante.

La tercera y última capa llamada capa de salida, realiza una combinación lineal de las funciones de base radial, es decir que las salidas proporcionadas por las neuronas de la capa oculta, son multiplicadas por el peso de su conexión y sumadas todas ellas para dar la salida de la red. Las neuronas de esta capa poseen activaciones lineales.

A diferencia del perceptron multicapa, este tipo de redes tiene carácter local, aunque dependiendo del tipo de aprendizaje podrían llegar a tener carácter global, por ejemplo en el método supervisado que se explicará más adelante.

Es este carácter local que caracterizan a este tipo de redes, lo que origina la rápida convergencia de las RNBR en relación a otras redes como el perceptron multicapa. Tanto las RNBR como el perceptron multicapa son aproximadores universales, utilizándose mucho en tareas de aproximación de funciones (interpolación), reconocimiento de patrones, memorias asociativas, etc.

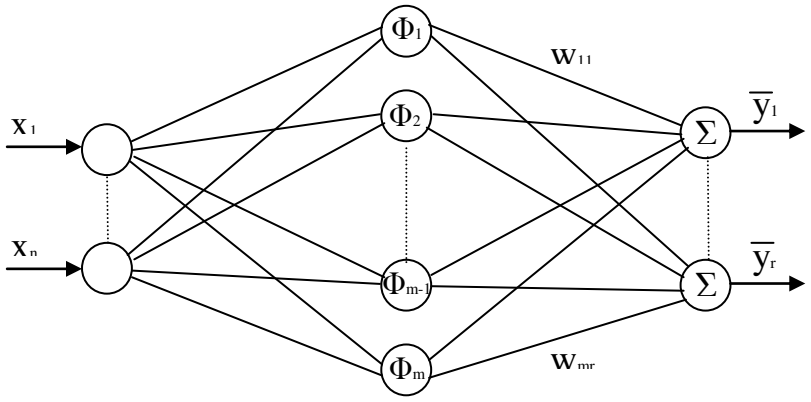


Figura 8: Estructura de una red neuronal de base radial

Como se puede ver en la figura 8, este tipo de redes tienen una capa de entrada, donde no se realiza ningún proceso, habiendo tantas neuronas como datos sean necesarios transmitir. La única capa oculta existente aplica una transformación no lineal sobre los datos de entrada, por medio de las funciones de base radial. Por último las salidas de la capa oculta se envían a las neuronas de la capa de salida, multiplicadas por el peso de la conexión, y sumadas todas ellas, en definitiva una combinación lineal de las salidas de la capa oculta.

Para una RNBR con *n* neuronas en la capa de entrada, *m* neuronas en la capa oculta, y *r* neuronas en la capa de salida, la fórmula que describe la salida de la red viene dada por la siguiente expresión:

$$\bar{y}_k(x) = w_0 + \sum_{i=1}^{m} w_{ik} \cdot \phi_i(x) \qquad k = 1,...,\mathrm{r}$$

donde w_{ik} son los pesos asociados a las conexiones de las neuronas *i* de la capa oculta con la neurona de salida *k*; w_0 es el umbral asociado a la neurona de salida, que en el caso de este simulador se ha supuesto siempre igual a 0; $x = (x_1,...,x_n)$ es la entrada

externa de la red, y $(\phi_i)_{i=1,...,m}$ son las funciones de base radial o funciones de activación de las neuronas de la capa oculta.

Las funciones de base radial se expresan normalmente como traslaciones de una función prototipo de la siguiente forma:

$$\phi_i(x) = \phi\left(\frac{\|x - c_i\|}{d_i}\right) \qquad i = 1,...,\text{m}$$

donde $c_i = (c_{i1},...,c_{in}) \in \Re^n$ es el centro *i* de la función ϕ_i, d_i es la desviación, anchura o factor de escala para el radio $\|x - c_i\|$, que es la distancia que existe entre el centroide *i* y el dato de entrada *x*; $\|\cdot\|$ es la norma euclídea definida como:

$$\|x - c_i\| = \left(\sum_{j=1}^{n} (x_j - c_{ij})^2\right)^{\frac{1}{2}} \qquad i = 1,...,\text{m}$$

La función ϕ que normalmente es una gaussiana, y es la que está implementada en el simulador, que es de la forma:

$$\phi(r) = e^{-\frac{r^2}{2}}$$

donde *r* es el radio o distancia que hay entre el centroide y el dato de entrada a la red neuronal.

La gaussiana suele ser la función más utilizada, pero se pueden utilizar cualquier tipo de función de base radial, de hecho el simulador implementa otras como la multicuadrática inversa, o la logistica. En la siguiente tabla se muestran algunas.

Nombre	Expresión	Parámetros
Gaussiana	$\phi(r) = e^{-\frac{r^2}{2\sigma^2}}$	Con parámetro de normalización $\sigma > 0$
Multi-Cuadráticas	$\phi(r) = (r^2 + \sigma^2)^{\frac{1}{2}}$	Con parámetro de normalización $\sigma > 0$
Multi-Cuadráticas Generalizadas	$\phi(r) = (r^2 + \sigma^2)^{\beta}$	Con parámetros $\sigma > 0$ y $1 > \beta > 0$
Multi-Cuadráticas Inversas	$\phi(r) = (r^2 + \sigma^2)^{-\frac{1}{2}}$	Con parámetro de normalización $\sigma > 0$
Multi-Cuadráticas Inversas Generalizadas	$\phi(r) = (r^2 + \sigma^2)^{-\beta}$	Con parámetros $\sigma > 0$ y $1 > \beta > 0$
Logística	$\phi(r) = \frac{2}{1 + e^{-\frac{r^2}{\sigma^2}}}$	Con parámetro de normalización $\sigma > 0$
Cúbica	$\phi(r) = r^3$	

Tabla 2: Ejemplo de funciones radiales

El principal problema de este tipo de redes, se produce con la alta dimensionalidad del espacio de entrada, es decir, cuando el número de neuronas en la capa de entrada es muy alto, se produce el problema llamado *curse of dimensionality.* Este problema provoca que el número de neuronas en la capa oculta, necesarias para obtener una buena generalización, crezca de forma exponencial.

Obviamente, y debido en gran parte a este problema, dependiendo del número de neuronas en la capa oculta se conseguirán resultados muy diferentes. La determinación de las neuronas necesarias para resolver un problema suele ser un factor crítico, aunque no el único, a la hora de conseguir una buena generalización de la red.

Normalmente la determinación de la topología de la red (nº de neuronas de la capa oculta), se hace de forma artesanal, es decir, el investigador utilizando su experiencia, determina el número de neuronas a utilizar, si bien el método de prueba y error es uno de los más utilizados cuando no se dispone de esta experiencia. En el caso de este trabajo, se han utilizado los algoritmos genéticos para encontrar un rango de parametrizaciones que haga que las aproximaciones conseguidas por la red de neuronas sean buenas.

En cualquier caso, los investigadores en este campo de la IA, han aportado sus soluciones, mediante las cuales se puede conseguir una aproximación razonable a lo que debería ser la topología óptima de la red. Si bien la mayoría de esfuerzos, dirigidos a la consecución de una forma automática de elección de la topología, han ido a parar a las redes perceptron multicapa, algunos investigadores han propuesto soluciones dirigidas a las RNBR.

Algunas de estas soluciones pasan por la utilización de un sistema multiagente [Valls, Galván & Molina, 00], donde hay un agente árbitro y varios agentes que se dedican al entrenamiento de la red con un número distinto de neuronas. Cada cierto número de ciclos los agentes de entrenamiento envían el error cometido al árbitro, el cual determina quién debe seguir con el entrenamiento (uno en cada momento) reduciendo así el número de ciclos total en la determinación del mejor agente.

Otro algoritmo es el algoritmo desarrollado por Platt [Platt, 1991], que se basa en ir añadiendo más neuronas a la topología y ver como va cambiando el error, añadiendo más neuronas a la capa oculta si se comprueba que el error disminuye.

4.2 - Entrenamiento de las RNBR

En esta sección se describe el entrenamiento de una RNBR, explicando paso por paso cada uno de los cálculos que hay que realizar.

Primeramente, cuando se crea una red neuronal, con una serie de neuronas en cada una de sus capas, hay que inicializar una serie de parámetros antes de entrenar la red. Primero se inicializan los centros de las neuronas también llamados centroides; debido a que el espacio de vectores de entrada se encuentra normalizado a valores pertenecientes al intervalo [0,1], los centroides se repartirán aleatoriamente en este intervalo.

Hay que indicar que los valores de entrada no pueden ser cualquier valor real, sino que estos datos necesitan primeramente de una normalización para poder ser tratados por la red neuronal. Los valores de salida de la red normalmente estarán normalizados también, devolviéndose valores que se encuentran dentro del intervalo [0,1].

Una vez que los centroides han sido posicionados dentro del espacio de vectores de entrada, el siguiente paso es inicializar las desviaciones de esos centroides, es decir la anchura que tendrá la gaussiana o la función de base radial utilizada. Las anchuras o desviaciones de los centroides, se calculan teniendo en cuenta los centroides anteriormente hallados. Esto es así porque se usa la distancia de los dos vecinos más cercanos a un centroide para calcular la desviación de ese centroide. La expresión utilizada en este caso es:

$$d_i = \sqrt{\alpha_{i1} \cdot \alpha_{i2}}$$

donde la desviación de la neurona *i*, es decir d_i, es la media geométrica de la distancia euclídea del centro c_i a sus dos vecinos más próximos. La implementación permite indicarle que sea cualquier número de vecinos, es decir que no tiene porque ser la distancia a los 2 vecinos más cercanos, sino por ejemplo la distancia a los n vecinos más cercanos, y también se ha implementado la máxima distancia a los n vecinos más cercanos, con el objetivo de realizar pruebas con el entrenamiento de la red.

Cualquier otro método se podría haber implementado, es la ventaja de implementar la red desde cero, como por ejemplo las distancias medias entre centroides, es decir calcular la distancia media que hay entre centroides, y aplicar esa distancia como desviación de todos los centroides, o cualquier otra que se nos pudiera ocurrir.

Una vez inicializados los centroides y sus desviaciones, el siguiente paso es la inicialización aleatoria de pesos también en el intervalo [0,1].

Una vez realizadas las inicializaciones de centros, desviaciones y pesos, se puede pasar al siguiente paso que es entrenar la red neuronal. Este entrenamiento se puede realizar de dos formas, de forma supervisada e híbrida, las cuales se explican de forma detallada en los siguientes apartados.

4.3 - Aprendizaje híbrido

Este tipo de aprendizaje entraña una primera fase en la que se determinan los centros y desviaciones de la red, de un modo no supervisado mediante el uso de determinados algoritmos, mientras que en la segunda fase se realiza la modificación de los pesos de la red de una forma, ahora sí, supervisada.

4.3.1 - Centros

La determinación de los centros de las neuronas se suele realizar, utilizando algún método de clasificación, como el algoritmo *K-Medias*.

Básicamente el algoritmo *K-Medias* realiza una clasificación del espacio de entrada, de forma que reparte este espacio de entrada entre los centroides de la red, dando lugar a regiones dominadas por uno u otro centroide, de tal forma que una región dominada por un determinado centroide, viene definido por el grupo de patrones de entrada más cercano al centroide que domina esa región.

Aparte del algoritmo *K-Medias*, para la determinación no supervisada de los centros, se pueden utilizar otros algoritmos basados en el agrupamiento de elementos del espacio de estrada, atendiendo a ciertas características de estos elementos. Por ejemplo, el algoritmo de *Kohonen* que realiza una clasificación de los puntos que se encuentran más cercanos en el espacio de entrada. Este algoritmo es utilizado en un tipo de redes neuronales llamadas *Mapas de Kohonen* [Kohonen, 2001], que es una red neuronal de 2 capas, que utiliza aprendizaje no supervisado, y está basado en redes competitivas.

Otro algoritmo podría ser el basado en el análisis de cluster estadístico que realiza también una clasificación de elementos, agrupándolos respecto a las características que poseen. La agrupación se realiza en base al concepto de distancia como la distancia Manhattan, la distancia de Chebychev o la de Minkowski que también han sido implementadas.

4.3.2 - Desviaciones

Una vez calculados los centroides, el siguiente paso es el cálculo de las desviaciones de cada uno de éstos, utilizándose la media geométrica de la distancia a los dos vecinos más cercanos, aunque no tiene porqué ser a los dos vecinos, podría ser a los tres o n vecinos más cercanos.

$$d_i = \sqrt{\alpha_{i1} \cdot \alpha_{i2}}$$

Cuando el número de neuronas de la capa oculta es menor que tres, obviamente no se puede utilizar esta expresión. Cuando sólo existe una neurona en la capa oculta, su desviación es $d = 1$, mientras que cuando hay dos neuronas solamente, la fórmula utilizada es $d = \alpha / 2$ para las dos neuronas siendo α la distancia que existe entre ambas neuronas.

Cuanto mayor sea la desviación asociada a una neurona, mayor será el campo de acción de esa neurona, de tal forma que una neurona da mayores valores de salida para un patrón que está cerca de esa neurona (centroide), y tanto más cuanto más grande es la desviación. Lo que ocurre es que por ejemplo, una neurona puede dar valores próximos

a cero para un determinado patrón (de ahí el carácter local de las neuronas), pero si se aumenta la anchura o desviación asociada a la neurona, entonces la salida para ese patrón puede variar enormemente.

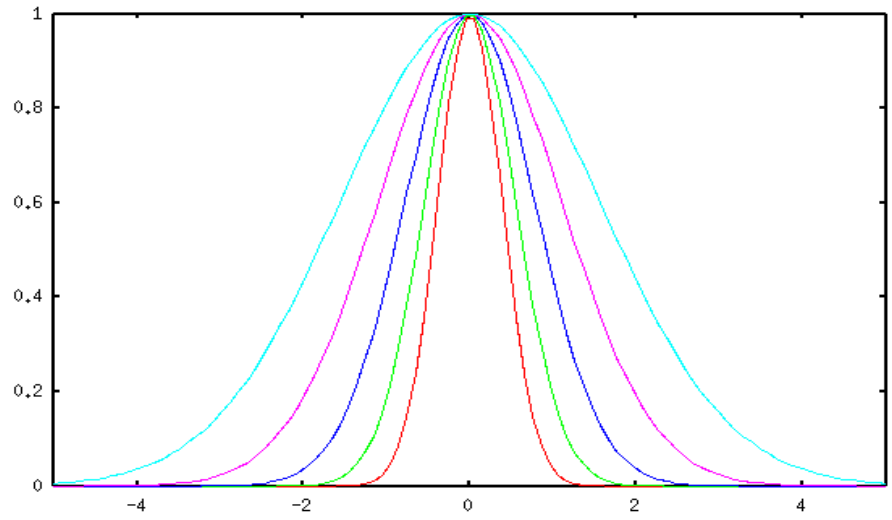


Figura 9: Gaussianas con distintas desviaciones

En la figura 9 se puede comprobar la influencia de la anchura (desviación) en la activación de una neurona; considerando que el eje de abscisas es la distancia de un patrón de entrada a una determinada neurona, y el eje de ordenadas es el valor devuelto por la neurona, se observa que mientras determinadas gaussianas devolverían valores próximos a cero para una distancia dada, para gaussianas con mayor anchura el valor devuelto puede ser bastante alto.

Las desviaciones pueden influir mucho en la capacidad de generalización de una red neuronal u otra. Para ver esto con mayor claridad, piénsese primero en obtener unos errores de aprendizaje lo más bajos posibles para lo cual lo más sencillo es tener tantas neuronas como patrones de entrada; si los valores de las anchuras son altos, las neuronas tendrán poco carácter local con lo que una neurona dará valores altos para patrones que estén alejados de ella, interfiriendo en la salida que sobre ese patrón de una neurona que esté más cerca de ese patrón, y aumentando el error de aprendizaje. Si las anchuras son pequeñas, posiblemente las salidas que las neuronas den para un patrón alejado serán próximas a 0, dando valores altos sólo para los patrones que estén más cerca de esa neurona y el error de aprendizaje baje.

El problema es que cuando se entrena la red hay que atender también a los errores de validación o test, y este error es el que determina la capacidad de generalización de la red, que no es más que la reacción de la red ante nuevos datos no aprendidos, es decir que para valores de entrada no aprendidos, la salida de la red ha de ser lo más próxima a la salida real. Si las desviaciones no son lo suficientemente grandes como para atender a esos nuevos patrones, las neuronas devolverán valores muy pequeños, y la salida de la red no se va a adecuar a los valores reales. En definitiva, entender que la desviación asignada a una neurona es algo muy importante, y que con desviaciones pequeñas, el error de entrenamiento puede ser pequeño pero el de test puede ser muy grande, con lo que hay que buscar que ambos errores sean parecidos y lo más bajos posibles.

4.3.3 - Pesos

Una vez que se han calculado los centros y desviaciones, el siguiente paso es el cálculo de los pesos, que se realiza de forma supervisada, por medio de un procedimiento de cálculos iterativos, basado en el método del gradiente descendente que trata de minimizar el error cuadrático medio, que se produce por la diferencia de los datos devueltos por la red, y la salida deseada, y cuya expresión viene dada por:

$$E = \frac{1}{Np}\sum_{N=1}^{Np} e_N \qquad \Longrightarrow \qquad e_N = \frac{1}{2}\sum_{k=1}^{r}\left(y_k(x_N) - \bar{y}_k(x_N)\right)^2$$

donde *Np* es el número de patrones de entrada e $y_k(x_N)$ e $\bar{y}_k(x_N)$ son la salida deseada y la salida *k* de la red para el patrón *N*.

La modificación de los pesos de la red, se realiza de acuerdo a la dirección negativa del gradiente (gradiente descendente) del error *E,* es decir que suponiendo que el error se pudiera representar en función de los pesos como una superficie, partiendo de un valor aleatorio de esa superficie $W^{(0)} \in \Re^{nw}$, desplazándose en dicha superficie siguiendo la dirección negativa del gradiente de *E* en el punto $W^{(N-1)}$, alcanzando así un nuevo punto $W^{(N)}$ (figura 6), que estará más próximo al mínimo de la función E que el anterior. El

proceso continúa hasta que se minimiza la función E lo cual sucede cuando $\partial E / \partial w = 0$. En este punto los pesos dejan de sufrir cambios importantes y el proceso de aprendizaje finaliza.

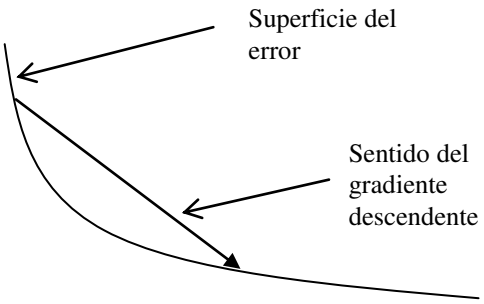


Figura 10: Gradiente descendente

Debido a la utilización del método del gradiente, los pesos se modifican de forma iterativa siguiendo la siguiente ley:

$$w_{ik}^{(N)} = w_{ik}^{(N-1)} + \alpha_1 \cdot \Delta_N w_{ik} \qquad \begin{matrix} i = 1,...,m \\ k = 1,...,r \end{matrix} \qquad \Delta_N w_{ik} = -\frac{\partial e_N}{\partial w_{ik}}$$

como se tiene que :

$$\bar{y}_k = \sum_{i=1}^{m} w_{ik} \phi_i(x) \qquad \frac{\partial e_N}{\partial w_{ik}} = (y_k - \bar{y}_k) \frac{\partial \bar{y}_k}{\partial w_{ik}} \qquad \frac{\partial \bar{y}_k}{\partial w_{ik}} = \phi_i(x)$$

la ecuación finalmente queda como:

$$w_{ik}^{(N)} = w_{ik}^{(N-1)} + \alpha_1 (y_k - \bar{y}_k) \phi_i(x)$$

de tal forma que α_1 es la razón de aprendizaje, y_k e $\bar{y}_k$ son la salida deseada y la salida de la red, respectivamente, mientras que ϕ_i es la función gaussiana aplicada a la neurona oculta *i*. Por tanto, el peso actual de la conexión que va de la neurona *i* de la capa oculta, a la neurona *k* de la capa de salida, es igual al anterior peso de esa misma conexión, más el producto de la razón de aprendizaje por la diferencia entre salida

deseada y salida de la red, por la salida de la neurona oculta *i* una vez aplicada la función de activación (en caso del simulador la función gaussiana), a un determinado patrón de entrada.

El método del gradiente descendente es un método indirecto, ya que se va iterando cada vez hasta alcanzar errores que prácticamente no cambian, en cuyo momento debería pararse el entrenamiento.

En el caso del aprendizaje híbrido, existe un método directo ya que la determinación de los pesos, es un problema de optimización lineal que se puede resolver algebraicamente mediante la resolución de la matriz pseudoinversa cuya solución son los pesos que minimizan el error.

En este caso la solución al problema de optimización viene dado por la expresión siguiente:

$$w = (M^t \cdot M)^{-1} \cdot M^t \cdot T$$

donde M y T son de la forma:

$$M = \begin{bmatrix} \phi_1(x_1).....\phi_m(x_1) \\ ..... \\ \phi_1(x_{Np}).....\phi_m(x_{Np}) \end{bmatrix} \qquad T = \begin{bmatrix} y_1(x_1).....y_r(x_1) \\ ..... \\ y_1(x_{Np}).....y_r(x_{Np}) \end{bmatrix}$$

donde los elementos de las filas de la matriz *M* son las salidas de cada neurona de la capa oculta, para cada uno de los patrones de entrada, y los elementos de la matriz *T* son las salidas deseadas para cada patrón de entrada. El elemento $\phi_i(x_1)$ representa la salida de la i-ésima neurona oculta para el primer patrón de entrada. Por tanto, la matriz tendrá tantas filas como número de patrones de entrada, y tantas columnas como neuronas ocultas.

4.4 - Aprendizaje supervisado

En este tipo de aprendizaje, no se utilizan algoritmos que calculen centros y desviaciones, sino que tanto los centros, como desviaciones y pesos, son calculados en la propia fase de entrenamiento, modificando sus valores atendiendo a los patrones de entrenamiento.

El entrenamiento en este tipo de aprendizaje se produce de manera supervisada, y es un proceso de cálculos iterativos basados en el método del gradiente descendente, que minimizan el error.

La expresión utilizada para modificar los pesos de las conexiones es la misma que en el caso del aprendizaje híbrido, ya que la parte de modificación de pesos se hacía de forma supervisada mediante la siguiente expresión:

$$w_{ik}^{(N)} = w_{ik}^{(N-1)} + \alpha_1 (y_k - \bar{y}_k)\phi_i(x) \qquad \begin{matrix} i = 1,...,m \\ k = 1,...,r \end{matrix}$$

Para los centroides la expresión es prácticamente la misma que para el caso de los pesos, ya que también se obtiene aplicando el método del gradiente descendente:

$$c_{ij}^{(N)} = c_{ij}^{(N-1)} + \alpha_2 \cdot \Delta_N c_{ij} \qquad \begin{matrix} i = 1,...,m \\ j = 1,...,n \end{matrix} \qquad \Delta_N c_{ij} = -\frac{\partial e_N}{\partial c_{ij}}$$

$$\bar{y}_k = \sum_{i=1}^{n} w_{ik}\phi_i(x) \quad \text{y} \quad \frac{\partial e_N}{\partial c_{ij}} = \sum_{k=1}^{r} (y_k - \bar{y}_k)\frac{\partial \bar{y}_k}{\partial c_{ij}} \Longrightarrow \frac{\partial \bar{y}_k}{\partial c_{ij}} = \phi_i(x) w_{ik} \frac{(x_j - c_{ij})}{d_i^{\,2}}$$

donde c_{ij} es la coordenada j del centroide i, x_j es la coordenada j del patrón de entrenamiento x, y d_i es la desviación asociada al centroide i. De este modo los centros se modifican siguiendo la siguiente ley:

$$c_{ij}^{(N)} = c_{ij}^{(N-1)} + \alpha_2 \sum_{k=1}^{r} (y_k - \bar{y}_k) w_{ik}\phi_i(x) \frac{(x_j - c_{ij})}{d_i^{\,2}}$$

La expresión para modificar las desviaciones sigue la misma filosofía que las anteriores. De este modo:

$$d_i^{(N)} = d_i^{(N-1)} + \alpha_3 \cdot \Delta_N d_i \qquad \Delta_N d_i = -\frac{\partial e_N}{\partial d_i}$$

$$\bar{y}_j = \sum_{i=1}^{n} w_i \phi_i \qquad \frac{\partial e_N}{\partial d_i} = \sum_{k=1}^{N} (y_k - \bar{y}_k) \frac{\partial \bar{y}_k}{\partial d_i} \qquad \frac{\partial \bar{y}_k}{\partial d_i} = \phi_i(x) w_{ik} \frac{\|x - c_i\|^2}{d_i^3}$$

finalmente:

$$d_i^{(N)} = d_i^{(N-1)} + \alpha_3 \sum_{k=1}^{r} (y_k - \bar{y}_k) \phi_i(x) w_{ik} \frac{\|x - c_i\|^2}{d_i^3}$$

donde el parámetro α_3 es la razón de aprendizaje, y podría ser distinta para modificar pesos, centros, pesos y desviaciones. En el simulador se utiliza la misma en las tres fórmulas (pesos, centros y desviaciones).

Como se ha comentado anteriormente, este tipo de entrenamiento pierde el carácter local que tiene el entrenamiento híbrido, ya que aquí los centros se van modificando y con ellos las desviaciones, las cuales podrían llegar a ser bastante grandes, afectando a grandes regiones del espacio de entrada.

4.5 - Algunas nociones sobre el aprendizaje

A la hora de entrenar una red neuronal, hay que tener en cuenta una serie de puntos, que pueden ayudar a que la red aprenda de forma conveniente. Es obvio que ciertos parámetros que se pueden modificar a la hora de entrenar una red, van a incidir de forma positiva o negativa en los resultados obtenidos.

También los patrones escogidos a la hora de entrenar y validar una red, pueden arrojar ciertos datos que determinen que éstos no son los más adecuados, y que deberían escogerse de otra forma.

A continuación se explican más en detalle estas cuestiones, y se dan ciertos consejos de los expertos en redes de neuronas para evitarlas.

4.5.1 - Sobreaprendizaje

Uno de los problemas que pueden aparecer cuando se entrena una red neuronal es el sobreaprendizaje. Este término se utiliza cuando una red es capaz de aprender de forma muy precisa los patrones con los que se entrenó, pero cuando se le presentan patrones que no intervinieron en el aprendizaje, también denominados patrones de test o validación, la red puede devolver valores muy alejados de la realidad.

Para que el usuario pueda ver si la red que está entrenando está sufriendo de sobreaprendizaje, es útil realizar la validación de la red cada cierto número de ciclos de entrenamiento.

La validación de la red no forma parte del aprendizaje, de tal forma que aunque se produzcan ciertos ciclos de validación mientras se está entrenando la red, no se modifican los pesos de acuerdo a estos patrones de validación por lo que no se aprenden estos patrones. El proceso de validación es sólo una forma de determinar si la red esta aprendiendo de forma apropiada o no, o lo que es lo mismo, si es capaz de generalizar o no.

Cuando se entrena la red y a la vez se valida, se pueden comparar los errores de entrenamiento y de validación, de tal forma que si el error de validación es mucho más grande que el de entrenamiento, es que la red no aprende bien, es capaz de aprender los patrones de entrenamiento pero no es capaz de generalizar con lo que el error de validación es grande.

Lo ideal, es que tanto el error de entrenamiento como el de validación sean lo más bajos posibles, y que ambos errores sean parecidos.

A veces el problema del sobreaprendizaje viene provocado por los patrones utilizados, de forma que los patrones utilizados para entrenar, son muy distintos a los usados para validar. En este caso se puede decir que no existe un sobreaprendizaje real, aunque sí una falta de generalización debido a una mala práctica en la elección de patrones.

Los patrones de validación siendo distintos a los de entrenamiento, deben parecerse en cierta medida a éstos. Supóngase que se desea que una red neuronal aprenda una función cualquiera, por ejemplo $y(x) = x^2$, y se tienen datos de un cierto intervalo de la función para utilizarlos como patrones de entrenamiento, supóngase que los valores normalizados de entrenamiento van de [0,1] con saltos de valor 0,001 entre cada valor consecutivo. Atendiendo a esto se tendrían 1001 patrones de entrenamiento, y se elegirían de aquí lo valores de validación por ejemplo 250 patrones. Desde luego, si los patrones están ordenados, sería un error elegir 250 patrones consecutivos (si el fichero está ordenado) como patrones de validación, ya que serían valores muy alejados de los utilizados para aprender. Lo ideal sería que se escogieran más o menos equiespaciados, de tal forma que si se escogen 250 patrones para validar, se escogiera un patrón de cada 4 de entrenamiento, para formar parte de los patrones de validación.

4.5.2 - Razón de aprendizaje

En el entrenamiento también influyen parámetros que hacen que el error obtenido por la red sea más grande o más pequeño, como es la razón de aprendizaje, que es un parámetro modificable por el usuario, y que va a determinar la medida del cambio de los pesos, así como de centros y desviaciones si es aprendizaje supervisado.

La modificación de los pesos de las neuronas, se realiza mediante una serie de saltos en la dirección del gradiente descendente del error como ya se ha indicado. Estos saltos serán más grandes o pequeños dependiendo de la razón de aprendizaje, y por tanto esto va a afectar también a los errores conseguidos.

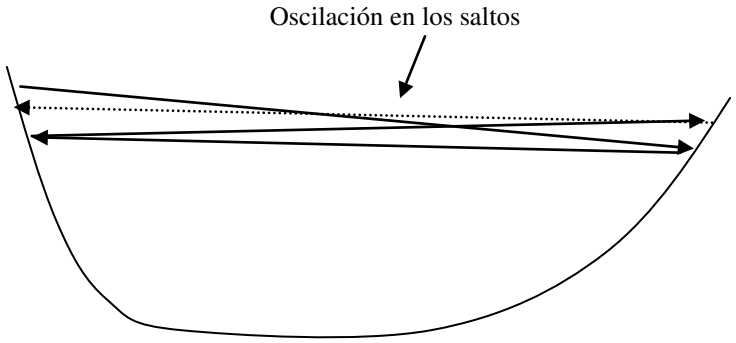


Figura 11: Implicación de la razón de aprendizaje en el error

En la figura 11 se ha representado la modificación de los pesos respecto de la superficie del error. Se ve como se baja hasta cierto punto en la dirección del mínimo error pero llega un momento en el que los saltos son tan grandes debido a la razón de aprendizaje, que no se consigue bajar ahí donde el error es mínimo.

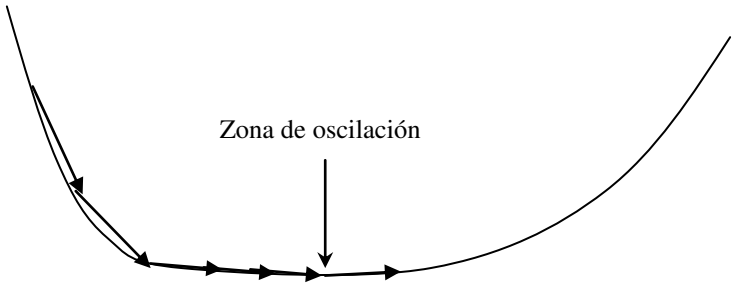


Figura 12: Implicación de la razón de aprendizaje en el error

En la figura 12 se ha utilizado una razón de aprendizaje más pequeña, de tal forma que se puede llegar a la zona más baja del error, ahí donde el error es mínimo. Se ve como los saltos permiten bajar mucho más que en el caso anterior, de tal forma que cuando el error vuelve a subir, el sentido del salto cambia porque se sigue el sentido del gradiente descendente, bajando de nuevo el error.

Hay que notar también que cuando la pendiente es muy pronunciada, lo cual se produce en los puntos donde existe una gran diferencia entre la salida de la red y la salida deseada, los saltos son mayores de acuerdo a la ley de modificación de los pesos (sección 3.4), ya que se multiplica la razón de aprendizaje por esa diferencia.

Hay casos, en los que una razón de aprendizaje alta ayuda a que el error conseguido sea más bajo. Esto, aunque resulte paradójico, puede darse, sólo hay que pensar en una superficie del error en los que existe algún mínimo local, de tal forma que nunca se consiga llegar al mínimo global, salvo porque el salto producido sea lo suficientemente grande como para saltarse el mínimo local, en el que quedaría estancado si los saltos fueran pequeños.

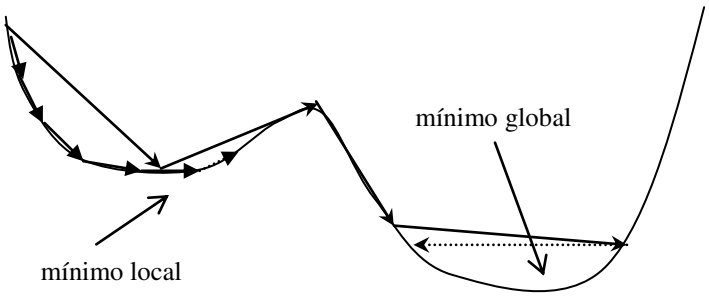


Figura 13: Ejemplo de mínimo local

En la figura 13 un mismo entrenamiento con una razón de aprendizaje más alta permite saltar el desnivel para llegar al mínimo global, mientras que el entrenamiento con una razón de aprendizaje más baja no lo permite, quedando estancados los valores de los pesos en un mínimo local.

El problema de los mínimos locales se produce sólo en el aprendizaje supervisado, ya que el problema a resolver es de tipo no lineal, mientras que en el caso del aprendizaje híbrido la determinación de los pesos es un problema lineal, y por tanto no cae en mínimos locales sino globales. Aunque en el aprendizaje híbrido la determinación de los pesos es un problema lineal, la determinación de los centros no es lineal, con lo que los pesos se adecuarán para conseguir el mejor valor para esa disposición de centros, si bien se podrían conseguir mejores resultados con otras disposiciones.

En general hay que utilizar razones de aprendizaje entre 0.01 y 0.25, según los expertos, pero es cierto que en las pruebas los mejores resultados se han conseguido con razones de aprendizaje que varían entre 0.3 y 0.8 o más.

4.5.3 - Elección de entrenamiento híbrido o supervisado

En general, no se puede decir que un tipo de entrenamiento sea mejor que otro a la hora de obtener mejores resultados.

Suele ser muy normal entrenar la RNBR utilizando el tipo de aprendizaje híbrido, y una vez que el entrenamiento se ha completado, realizar una serie de ciclos con entrenamiento supervisado.

Es conveniente realizar el nuevo entrenamiento con una razón de aprendizaje pequeña (por ejemplo $\alpha = 0.01$), sólo para refinar la red y obtener mejores resultados. Si se utilizara una razón de aprendizaje alta sería como entrenar desde el principio la red en modo supervisado, ya que tanto centros, como pesos y desviaciones cambiarían completamente, y los datos conseguidos en el entrenamiento híbrido no habrían servido de nada.

El entrenamiento híbrido tiene la característica de ser un entrenamiento mucho más rápido, por esta razón es el elegido cuando se desea implementar una RNBR que funcione en tiempo real, ya que el entrenamiento supervisado ha de cambiar centros, pesos y desviaciones, mientras que el híbrido tan solo modifica los pesos.

Sólo la experimentación determinará el tipo de aprendizaje que es mejor para un problema dado, teniendo en cuenta que el aprendizaje supervisado rompe un poco con el esquema de las RNBR's ya que elimina el concepto de la localidad de las neuronas, generalmente recomendándose su uso sólo para refinar los resultados obtenidos con el aprendizaje híbrido.

En el caso que nos ocupa los mejores resultados se han obtenido con el aprendizaje totalmente supervisado, y no con el híbrido, quizá sea debido a que el tipo de problema fuera una serie temporal, el caso es que siempre los mejores resultados se han obtenido con este tipo de aprendizaje, mientras que cuando la red se usaba para interpolar por ejemplo e^x se obtenían mejores resultados y con muchos menos ciclos de aprendizaje, con el entrenamiento de tipo híbrido.

4.6 - Principales características de las RNBR's

Las RNBR´s tienen una serie de características que se podrían resumir en las siguientes:

- Las RNBR´s son aproximadores universales de funciones.

- Surgen como solución regularizada al problema de interpolación.

- El entrenamiento es rápido pero pueden necesitar más neuronas que otros tipos de redes.

- Emplean funciones radiales, normalmente gaussiana como función de activación. otorgando propiedades locales a las redes.

- Son redes de tres capas (entrada, oculta y salida) de tipo feedforward.

- Dos tipos de entrenamiento, híbrido o totalmente supervisado, con distintas características.

- La determinación de los pesos de la red mediante el método híbrido, es un proceso de optimización lineal que se reduce a moverse por la superficie del error hasta llegar al mínimo global, en el caso de utilizar el método indirecto del gradiente descendente, o a operar con matrices hallando la matriz pseudoinversa, si se utiliza el método directo.

- Utilización de métodos de clasificación para la determinación de centros, y métodos heurísticos para el cálculo de las desviaciones (método de los n vecinos más cercanos).

CAPITULO 5 – ALGORITMOS GENÉTICOS

5. ALGORITMOS GENÉTICOS

5.1. INTRODUCCIÓN A LOS ALGORITMOS GENÉTICOS

5.2. FUNCIONAMIENTO DE LOS ALGORITMOS GENÉTICOS

CAPÍTULO 5 – ALGORITMOS GENÉTICOS

En este capítulo se expone una introducción a los algoritmos genéticos, hacía qué tipo de problemas van dirigidos y cómo se relacionan con la biología.

También se explicará cómo funcionan estos algoritmos y la forma de modelar un problema para ser resuelto mediante este tipo de algoritmos, así como las cosas que hay que tener en cuenta si se desea implementar uno de ellos.

5.1 – Introducción a los Algoritmos Genéticos

Los algoritmos genéticos forman parte de la computación evolutiva, la cual se engloba dentro del aprendizaje de tipo supervisado como las redes de neuronas, aunque las redes de neuronas también admiten el aprendizaje no supervisado.

Como en el caso de las redes de neuronas se basa en lo observado en la naturaleza, y está inspirada en la teoría de la evolución de Darwin.

Dentro de la computación evolutiva se pueden encontrar los algoritmos genéticos que son los que se han programado, las estrategias evolutivas, la programación genética o los sistemas clasificadores.

Los algoritmos genéticos extrapolan ideas del mundo real como se ha comentado, donde por ejemplo la representación del problema sería una cadena de datos, cuya extrapolación sería la cadena de ADN biológica que lleva codificado un individuo de determinada especie.

El funcionamiento de los algoritmos genéticos, como reflejo de la naturaleza es muy fácil de entender. Se parte de una población inicial donde cada individuo tiene sus características que le hacen único, esto es lo que viene codificado en la cadena solución.

La forma de codificar las cadenas de la población inicial es de forma aleatoria, tenemos un problema que se quiere resolver, pero se desconoce el modo de determinar la parametrización óptima de la solución, por lo tanto cada individuo creado con cadenas aleatorias son posibles soluciones del problema.

El siguiente paso es la evaluación de cada individuo, por lo que es necesario disponer de una forma de determinar si cierto individuo es mejor que otro, si esto no se puede determinar no se pueden usar este tipo de algoritmos. La evaluación se correspondería con la función de evaluación que se verá más adelante, y sería el valor que se muestra por parte del cromosoma, siendo el fenotipo el valor real de la cadena del cromosoma.

Una vez evaluados cada uno de los individuos de la población inicial, se inicia la etapa de sobrecruzamiento, donde los mejores individuos se cruzan entre sí para dar lugar a nuevos individuos. En esta etapa también aparece o no lo que se denomina mutación, y que puede afectar a las cadenas de los individuos generados por cruzamiento, o de cualquier otro individuo.

De esta forma se van generando nuevas poblaciones de individuos por sobrecruzamiento y mutación, y una proporción de cada población, por ejemplo el 20% de los mejores individuos persisten a la siguiente generación, evaluándose para cada nueva población cuales son los mejores individuos, desapareciendo paulatinamente los peores, realizándose de esta manera un proceso de búsqueda que lleva poco a poco a ir optimizando la solución.

Los algoritmos genéticos no aseguran hallar el óptimo global, pero sí que con ellos se puede llegar a soluciones muy buenas.

La computación evolutiva se utiliza en general para encontrar soluciones a problemas de optimización complejos, de los cuales no se conoce una técnica directa a partir de la cual se pueda obtener la solución. Estos algoritmos se han usado en problemas tan dispares como los relativos a optimización combinatoria y numérica, problemas de planificación y control, problemas de diseño, minería de datos, aprendizaje automático, etc.

Las ventajas de la computación evolutiva son que requieren poca información del dominio del problema, son algoritmos fáciles de implementar, funcionan como solucionadores de propósito general, y además se pueden mezclar con otras técnicas de resolución de problemas.

Entre las ventajas de estos algoritmos se encuentran la facilidad de la interpretación de las soluciones propuestas, permiten la ejecución interactiva, y además producen tantas soluciones alternativas como se le indique.

Otras características de este tipo de algoritmos es que los resultados que produce son aceptables a un costo computacional bajo. Si el tiempo de cómputo que se le permite en ejecución es más alto, los resultados serán mejores.

Los algoritmos genéticos tienen paralelismo implícito efectúan búsqueda robustas, esto es porque estos algoritmos muestrean de forma implícita hiperplanos, de tal forma que los individuos codificados de forma binaria constituyen un vértice de un hipercubo que viene a representar el espacio de búsqueda, siendo miembro de $2^L - 1$ hiperplanos distintos, siendo L la longitud de la cadena binaria que codifica el cromosoma del individuo. Por otro lado existen $3^L - 1$ hiperplanos del espacio de búsqueda total. De esta forma el individuo no provee de mucha información por separado, pero la búsqueda en una población es más que la suma de los individuos, de forma que existen muchas competiciones entre hiperplanos en una población. El muestreo de muchos hiperplanos en la evaluación de ristras es lo que se conoce como paralelismo implícito.

Este tipo de algoritmos cuentan con otra serie de ventajas sobre otros, y es que se pueden aplicar a problemas muy complejos, donde los datos con lo que se cuente sean incompletos, o cuyas relaciones entre los parámetros sean muy complejas. Se aplican a problemas donde el espacio de búsqueda es muy grande, donde suelen existir bastantes óptimos locales. En relación a esto último, lo que se quiere indicar, es que con suficiente tiempo de cómputo, generando un número de poblaciones razonablemente alto, la calidad de la solución es buena, y aunque no encuentre el óptimo global, encontrará un buen óptimo local.

El espacio de búsqueda del problema en concreto, viene determinado por el conjunto de todas las posibles y distintas soluciones (individuos) que se puedan generar, donde una posible medida de la complejidad del problema podría ser el del tamaño del espacio de búsqueda. Obviamente esto último no quiere decir que se deban generar todos esos individuos, el tiempo de cómputo a día de hoy, para resolver ciertos problemas con espacios de búsqueda inmensos, sería de años o más, pero en el caso en el que se pudieran generar todos los posibles individuos (soluciones) del espacio de búsqueda, es claro que se llegaría al óptimo global, pero en ese caso no sería necesario el uso de la computación evolutiva.

La aplicación de los operadores de sobrecruzamiento y mutación, implican una búsqueda pseudo-aleatoria por el espacio de búsqueda, ya que estos operadores no son deterministas, y esta búsqueda es una búsqueda dirigida, dirigida por la función de adecuación que determina qué individuos de la población son los mejores.

Mientras que el operador de sobrecruzamiento induce una búsqueda local por el espacio de estados, el de mutación implica una búsqueda global. Esto es así, porque como se ha comentado anteriormente, el sobrecruzamiento implica, que dos individuos con buena evaluación, al sobrecruzarse van a generar otro individuo muy parecido a los dos de los que parte, por eso se habla de una búsqueda local con este operador, mientras que el operador de mutación, puede generar individuos con su cadena de adn muy distinta a la inicial.

5.2 - Funcionamiento de los Algoritmos Genéticos

A la hora de utilizar un algoritmo genético son necesarias varias fases, la primera sería la modelización del problema a resolver. Esto implica que hay que diseñar como se va representar la solución al problema.

Una parte crucial, es contar con una función de adecuación, también llamada rutina de evaluación, mediante la cual se va a poder determinar qué individuos son mejores que otros, sin contar con esta función de evaluación es imposible utilizar un algoritmo genético, porque no se conocería cómo de cerca de la solución está cada individuo, por lo que no se podría elegir ninguno entre ellos.

Lo siguiente a determinar es el tamaño de cada población, pocos individuos acotarán demasiado el espacio de búsqueda, de tal forma que será difícil llegar a buenas soluciones, pero poblaciones muy elevadas inducirán un mayor tiempo de cómputo, aunque también un espacio de búsqueda mayor, hay que buscar un compromiso por tanto entre tiempo de cómputo y espacio de búsqueda.

Otro factor que determina si el espacio de búsqueda va a ser más o menos amplio es el operador de mutación, que como se ha comentado es el encargado de globalizar la búsqueda. Hay pues que determinar qué tasas de mutaciones es la apropiada, teniendo en cuenta que una tasa alta de mutación, puede hacer que en la búsqueda se den demasiados bandazos de un lado a otro, sin refinar demasiado la solución al problema.

Hay que considerar también el tipo de sobrecruzamiento, valorando qué cantidad de los individuos de la población pueden sobrecruzarse, por tanto cuantos individuos pueden aparecer en la nueva población. Si se permiten muchos individuos, al final cada población será más endogámica, por lo que hay que limitar el número de de individuos sobrecruzados de cada población, para que el espacio de búsqueda no sea tan limitado.

Por último, hay que determinar cuando se debe terminar la búsqueda, por ejemplo cuando la evaluación del individuo encontrado en cierto momento sea superior o inferior a cierto valor marcado, o bien cuando el tiempo de cómputo sobrepase cierto

valor, o cuando el número de poblaciones generadas, sea mayor que un valor determinado. En la práctica suele ser un compendio de los anteriores factores, es decir, que se puede pensar en parar cuando se encuentre un individuo con cierto resultado en su evaluación, por encima o por debajo a uno dado, pero que si bien el número de poblaciones o el de cómputo no se han sobrepasado que se continúe buscando.

Haciendo una comparación con el mundo biológico, se debe encontrar la forma de representar un cromosoma, el cual será una posible solución al problema, mediante una estructura, por ejemplo una cadena de bits, que sea una posible parametrización del problema, esta cadena puede contener enteros, reales, caracteres, enumerados, etc, en definitiva cualquier tipo de datos del lenguaje en el que se programe.

Por lo tanto los pasos serían, primero generar una población con cromosomas aleatorios, que sean posibles soluciones al problema propuesto. A continuación a esa población se le aplica la función de evaluación, donde se determina cuales son los mejores individuos.

La siguiente fase es aplicar los operadores de mutación y sobrecruzamiento de acuerdo al operador de selección sobre la población generada, teniendo en cuenta los mejores individuos, generando una nueva población, donde persisten los mejores individuos más otros que podrían ser mejor solución al problema.

Por último los individuos generados reemplazarán a los peores individuos de la población, generando una población nueva, a la que se le aplicará de nuevo la función de evaluación pasando por todos los pasos anteriores. Terminando cuando la condición de salida se cumpla.

El operador de selección que determina qué individuos se van a sobrecruzar, se puede implementar de diversas maneras, por ejemplo mediante una probabilidad proporcional al valor de adecuación, con lo que es posible que individuos con valores bajos de evaluación puedan dejar descendientes, y así expandir el espacio de búsqueda.

También se puede realizar por rango, donde aquellos en los primeros puestos se pueden sobrecruzar hasta completar el porcentaje de sobrecruzados de la nueva población, de esta forma los que están posicionados más abajo en el ranking no se podrán sobrecruzar.

Otro método es la selección por torneos, que se puede ver como una forma particular de la selección por rangos. En este caso se selecciona un conjunto de individuos de la población de manera aleatoria, y luego se escoge el individuo con mejor evaluación. Luego se vuelve a escoger de forma aleatoria otro rango de individuos y así, de forma que los escogidos son los que se pueden sobrecruzar hasta cumplir con el porcentaje de sobrecruzamientos de la nueva población.

Normalmente la formula usada suele ser la de selección por probabilidad proporcional, este método tiene el problema de que enseguida se produce una convergencia en la población, de forma que todos los individuos cuentan con unos cromosomas parecidos, con unos valores de adecuación parecidos entre ellos, lo que limita la búsqueda.

En la selección por rango, se dice que no hay una justificación biológica para ella, aunque evita la convergencia prematura de la población. Por su parte la selección por torneos si encuentra una justificación biológica con las ventajas de la selección por rango.

El operador de sobrecruzamiento por su parte, tiene su contrapartida biológica siendo la forma más común de recombinación. En este caso el sobrecruzamiento se realiza con una probabilidad p, siendo 1-p la probabilidad de que no haya sobrecruzamiento, en cuyo caso se aplica la clonación al individuo.

La forma de dividir dos individuos para generar otro nuevo es mediante el método de los n puntos, de tal forma que se cogen dos cromosomas y se dividen por n puntos para dar como resultado n+1 segmentos. El corte es el mismo en los cromosomas, abajo se pueden ver lo que sería n=2 puntos y 3 segmentos para dos cromosomas distintos.

1	0	0	1	1	1	0	1	1	0

0	1	1	0	1	1	0	0	1	1

Hay que tener en cuenta que el corte no puede ser arbitrario, hay que tener en cuenta que los parámetros tienen una longitud y hay que respetarla, lo que sí se puede hacer es desde realizar tantos cortes como parámetros representen al cromosoma, a realizar agrupaciones, recogiendo grupos de parámetros en cada corte.

Una vez realizado los n cortes, en el ejemplo 2 cortes, se conforma un nuevo individuo con la selección de cada uno se los segmentos alternativamente de cada progenitor, resultando en el caso del ejemplo, un nuevo individuo con el cromosoma que aparece más abajo.

1	0	1	0	1	1	0	1	1	0

Como se puede ver, se han escogido el primer y último segmento (gen) del primer progenitor, y el central del segundo progenitor, conformando de esta forma un nuevo individuo, para la siguiente población que tendrá que ser evaluado.

Como ya se ha explicado, el operador de mutación permite la exploración de áreas de búsqueda más amplias, que con solo operador de sobrecruzamiento no se podrían alcanzar. Este operador aporta nueva información a los individuos, previniendo la convergencia de cromosomas en la población, y permitiendo llegar a soluciones nuevas y mejores.

Esta tasa de mutación es baja, ya que de otra manera la búsqueda por el espacio de estados se podría desorientar, dando bandazos de un lado a otro sin encontrar buenas soluciones, pero si es muy baja la búsqueda se puede estancar produciéndose la convergencia de los cromosomas en un corto lapso de tiempo.

Hay ciertos estudios en cuanto a cuál debería ser la tasa de mutación idónea, por ejemplo *De Jong* proponía usar $P_m = \frac{1}{L}$ [De Jong, 1975] siendo P_m la probabilidad de mutación, y L la longitud del cromosoma. Otros investigadores recommiendan que esta tasa de mutación sea $P_m = (M * L^{\frac{1}{2}})^{-1}$ [Hesser, J., y Männer, R,1991], donde M es el tamaño de la población.

Desde el punto de vista de la renovación de la población en cada generación, existen dos aproximaciones, una propone el reemplazo total, y la segunda un reemplazo parcial, dejando que los mejores individuos puedan persistir varias generaciones, hasta ser reemplazados por otros mejores.

Cuando el reemplazo es parcial existen distintos criterios, como sólo reemplazar a los progenitores, o a los individuos con peor evaluación, o la eliminación de los individuos más parecidos. El primer y último criterio difuminan de alguna manera el problema de la convergencia, permitiendo una mayor disparidad de cromosomas en la población.

CAPITULO 6 – EXPERIMENTOS Y RESULTADOS

6. EXPERIMENTOS Y RESULTADOS

6.1. EXPERIMENTOS

6.2. RESULTADOS

6.3. CONCLUSIONES

6.4. TRABAJOS FUTUROS

CAPÍTULO 6 – EXPERIMENTOS Y RESULTADOS

En este capítulo, se tratarán los aspectos relativos a la obtención de resultados, cómo se ha parametrizado la red de neuronas y por qué se ha hecho así.

Se verán los resultados de cada uno de los experimentos de predicción de la probabilidad de muerte anual, tanto por tabla de mortalidad de cohorte, como por tabla de mortalidad por periodo.

Por último se dará una explicación a los resultados obtenidos, y se indicará cómo se piensa que se podrían mejorar.

6.1 – Experimentos

Como ya se ha explicado hay muchos grados de libertad en una red de neuronas de base radial, no sólo hay que posicionar cada neurona en unas coordenadas del espacio de datos, sino que también hay que elegir la función de activación, y seleccionar las desviaciones para cada neurona.

También entraban en juego el tipo de entrenamiento, si era híbrido, totalmente supervisado, o una mezcla de ambos, por otro lado si en caso de ser entrenamiento híbrido se iba a usar el algoritmo K-Medias con tipo de distancia Euclídea, Manhattan, Minkowski, Chevychev etc.

Cuál iba a ser el número de neuronas de entrada, y cuál el de la capa oculta, ya que de salida se había determinado que iba a ser sólo una, ya que proporcionar varias salidas de la red no iba a redundar en principio en una gran precisión, por lo que esto se fijo desde el principio.

Pues bien, una vez implementada la red de neuronas había que probar su funcionamiento, y lo más sencillo sería ver como aproximaba una función, en concreto

se escogió la de la figura de abajo, que se corresponde con y=$\frac{1.5*(10x)^2}{e^{10x}}$ consiguiéndose una precisión muy buena con un error cuadrático medio del orden de 10^{-10}.

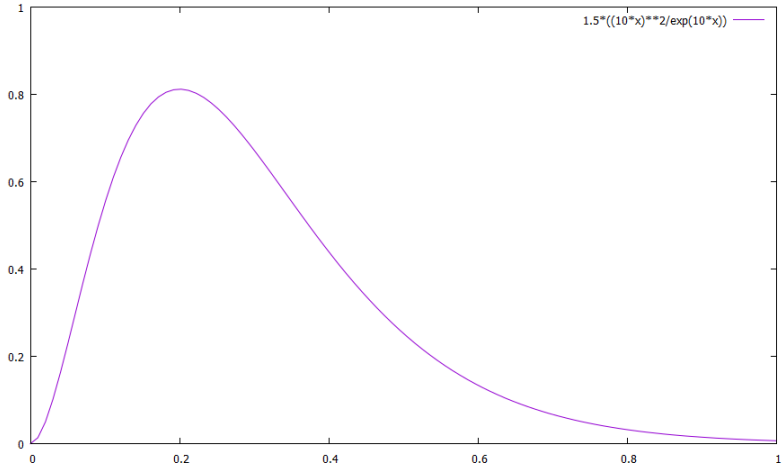


Figura 14: Función exponencial entre [0,1]

La forma de entrenar la red fue escogiendo un número de puntos determinado dentro del rango de [0,1], se escogieron igualmente espaciados, donde a su vez se escogían otros valores con los cuales no se entrenaba la red. En el momento en el que el error de entrenamiento empezaba a subir con respecto del error en el ciclo anterior se paraba el entrenamiento.

La modelización de este problema es muy sencilla, y cuenta con una entrada, n neuronas ocultas y una neurona de salida. La entrada es cualquier valor que tome la *x* de la función propuesta, mientras que la salida es el valor que toma *y*.

En general para este problema lo que mejor funcionó fue el entrenamiento híbrido que con uno totalmente supervisado, con unas 20 neuronas en la capa oculta, para 1000 patrones de entrenamiento y una razón de aprendizaje de 0,001. Pero el problema de la mortalidad no era un problema de interpolación, era una serie temporal y este es un problema un poco más complejo.

Modelizar el problema como una serie temporal, entraña que una o varias entradas a la red, dan como salida el momento siguiente a pronosticar. Supóngase que x_1 es el valor de la q_x en el momento 1 y así sucesivamente. Supóngase también que se escogen 4

instantes de tiempo para predecir el siguiente, por lo que las entradas estarían en verde y la salida deseada en azul.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}

Como se puede observar con 10 datos temporales, y cuatro entradas a la red, se dispondrían de 6 patrones de entrenamiento, donde las entradas a la red en verde, deberían proporcionar la salida marcada en azul, si es pronóstico a un año.
En el caso de que se quisiera pronosticar a 3 años por ejemplo, las entradas serían las mismas, pero la salida sería la marcada en naranja, pero en este caso se contaría como máximo con 4 patrones de entrenamiento. También se puede dar el caso que por cualquier razón se quiera pronosticar el elemento que está a uno y tres años.

Este entrenamiento se realiza hasta un determinado número de ciclos de entrenamiento, o hasta que el error de validación empiece a subir. En este caso la validación ha sido en relación a los 4 años anteriores, es decir, que de acuerdo al ejemplo de la tabla anterior, de acuerdo a los valores en x_7, x_8, x_9 y x_{10} para el caso de pronóstico a un año, considerando que los últimos valores son los más representativos de la serie, y son contra los que se debe validar. Es lógico pensar que el valor siguiente en la serie temporal, al menos en el caso de la mortalidad, será parecido a los n patrones anteriores, siendo n un valor arbitrario, pero que no debería ser muy alto para no desvirtuar la salida.

La experimentación comienza, y se descubre que los valores con aprendizaje híbrido que tan buenos resultados dio anteriormente, ya no son tan buenos, se modifican neuronas, razón de aprendizaje pero siguen siendo bastante decepcionantes.

En este punto entran los algoritmos genéticos a la palestra, donde los parámetros que se pueden modificar para cada cromosoma de la población son, el número de entradas de la red, el de neuronas ocultas, la salida se queda fija a una salida nada más por lo que no es parte del cromosoma, el tipo de entrenamiento, la razón de aprendizaje, la función de activación, el tipo de distancia para calcular los centroides por K-Medias y el cálculo de las desviaciones si por máxima distancia o de los 2 vecinos más cercanos.

Se lanza el experimento y se comprueba que los mejores individuos son los que han realizado un entrenamiento de tipo supervisado, frente a los de aprendizaje híbrido, por lo que se fija el tipo de entrenamiento a totalmente supervisado y se realiza otra búsqueda limitando eso sí el número de individuos revisados totales a 1000. Se comprueba como la razón de aprendizaje fluctúa entre 0,3 y 0,8 en los mejores individuos, por lo que la razón que se estaba usando en un principio de 0,01 resultaba por lo visto insuficiente, aunque en las primeras pruebas con la función exponencial dio unos resultados muy buenos.

La siguiente parte a probar era el número de entradas para el entrenamiento, en este caso hubo bastante disparidad, aunque en general los mejores tenían menos de 10 entradas, por lo que se limitó a 10 como máximo.

Después de estos datos, lo que se implementó es batería de generación de redes. Por ejemplo, se quería pronosticar a un año para la edad de 50, en modo cohorte, durante los próximos 5 años, es decir pronosticar las mortalidades a las edades de 51, 52, 53, 54 y 55 años, entonces se generaban redes que tenían desde una entrada hasta 10, con entrenamiento totalmente supervisado de hasta 3000 ciclos, con neuronas en la capa oculta que fluctuaban entre dos neuronas a 10 neuronas como máximo, con razones de aprendizaje que iban de 0.1 a 1, y para cada una de ellas 10 pruebas, por lo que al final se generan 10*9*10*10=9000 redes distintas, de las cuales la elegida era la que menor error de validación había tenido.

A partir de la red seleccionada, como se disponía de los datos reales para realizar el testeo, se introducía la entrada a la red, la cual producía una salida y se comparaba con la salida real que debía dar. En este punto se había pasado de una precisión del 15% al 20% en el mejor de los casos a un 98% o más, lo cual se le agradeció a los algoritmos

genéticos, que proporcionó un rango de parámetros entre los que se generaron redes de neuronas, que devolvían muy buenos valores.

En principio, el que los valores obtenidos a partir del tipo de red anterior fueran buenos, daban una idea de que la red había sido capaz de generalizar muy bien, pero claro, para que diera esas salidas se debían tener los datos reales de entrada. Por ejemplo, para pronóstico a un año y con una red de tres entradas, se le darían los 3 años anteriores al de pronóstico de entre los datos reales de la tabla, y la red proporcionaría el pronóstico del siguiente año. Para pronosticar el segundo año, se introduciría como entrada los tres años anteriores, dentro de los cuales no estaría el pronosticado sino el valor real de la tabla.

Lo anterior no resulta demasiado natural, porque si bien realiza la tarea de pronóstico a un año correctamente, si se desea pronosticar a más años, lo natural es hacerlo a partir de la salida de la red de neuronas, de tal forma y volviendo al ejemplo anterior, el pronóstico del segundo año se haría basándose en el pronóstico del primer año, que se le pasaría como entrada, en vez del valor real de la tabla como en el caso anterior.

En este último caso la red es una red con retroalimentación, de esta forma cuando se vean en las figuras de los pronósticos "*Con retroalimentación*" o "*Sin retroalimentación*", se puede entender a qué viene referida esta denominación y cuál es el funcionamiento de la red.

En general el error de una red con retroalimentación es mayor que una sin ella, por el simple hecho de que la primera considera como valores reales la salida de la red, acumulando más error en cada pronóstico. Sí que es cierto que en determinados casos, el estar retroalimentada o no, no da una variación muy grande en los resultados de salida, lo que no quiere decir que el error respecto a los datos reales sea más o menos elevado.

6.2 – Resultados

En este apartado se verán los resultados aportados por las redes consideradas mejores, de acuerdo a su error de validación.

En la figura se puede ver un pronóstico de tipo cohorte, siendo el periodo de entrenamiento que va del año 1960 al 2000, y pronóstico a un año, es decir para los 51 años en el 2001, 52, en el 2002, hasta llegar a los 55 en 2005. Lo que aparece en verde es la salida calculada por la red para cada uno de esas edades, y en morado los datos que aparecen en la tabla de mortalidad para esas edades. Se puede ver que está sobreestimando la mortalidad. Este caso es de una red no retroalimentada.

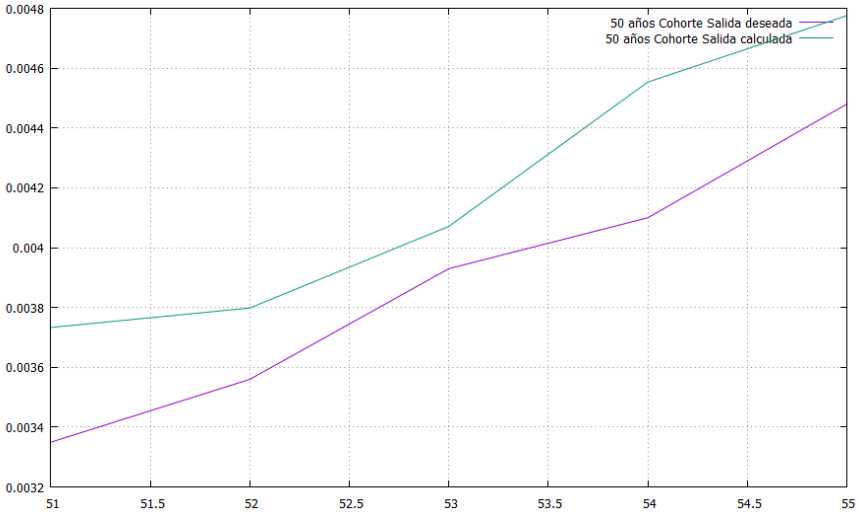


Figura 15: 50 años, Cohorte, pronóstico a un año

En la siguiente figura se puede observar el mismo caso, pero aquí la red es retroalimentada por sus propias salidas, mientras que en el caso anterior funcionaba con los datos reales de la tabla de mortalidad.

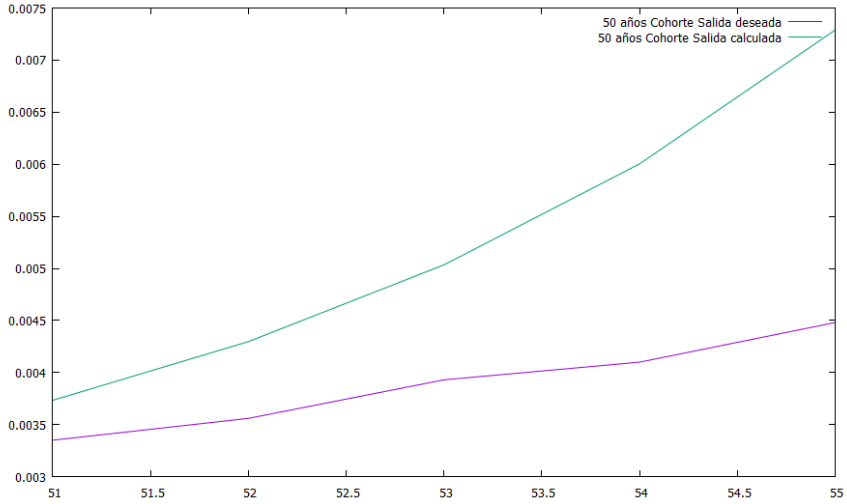


Figura 16: 50 años, Cohorte, pronóstico a un año retroalimentada

Se puede observar en este caso que la salida calculada difiere más de la real que en el caso anterior, y que en este caso la realimentación introduce un mayor error en la predicción.

Como se puede observar en la siguiente imagen al comparar la precisión en la estimación, ambas sobreestiman la mortalidad pero la retroalimentada tiene una menor precisión.

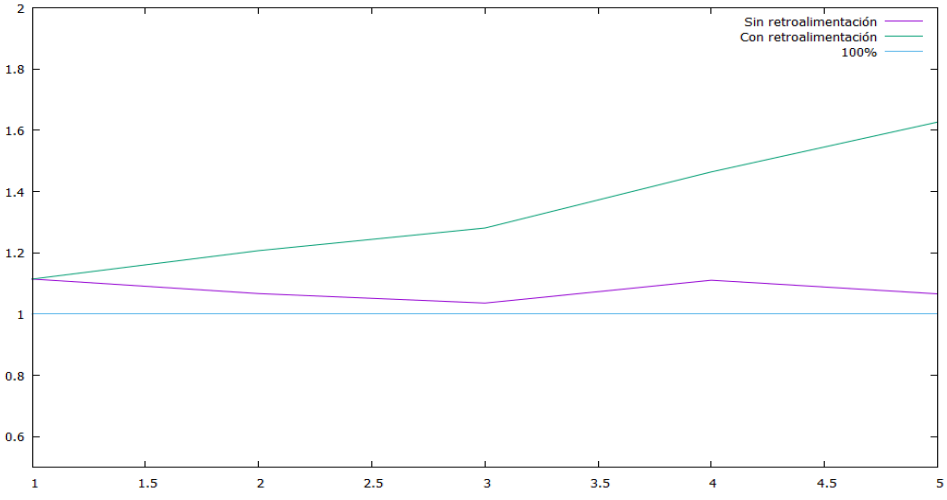


Figura 17: 50 años, Precisión entre ambos tipos de red

Ahora se han realizado los mismos experimentos que en el caso anterior, pero en este caso el pronóstico es a dos años, es decir que los pronósticos son para el 2002 con 52 años, para el 2004 con 54 años así hasta el 2010 con 60 años. Arriba a la izquierda se encuentra la previsión realizada por la red sin retroalimentar, a la derecha los resultados de la retroalimentada, y abajo la diferencia de precisión entre una y otra, y se puede observar que la precisión va decayendo, y que la red retroalimentada acumula más error.

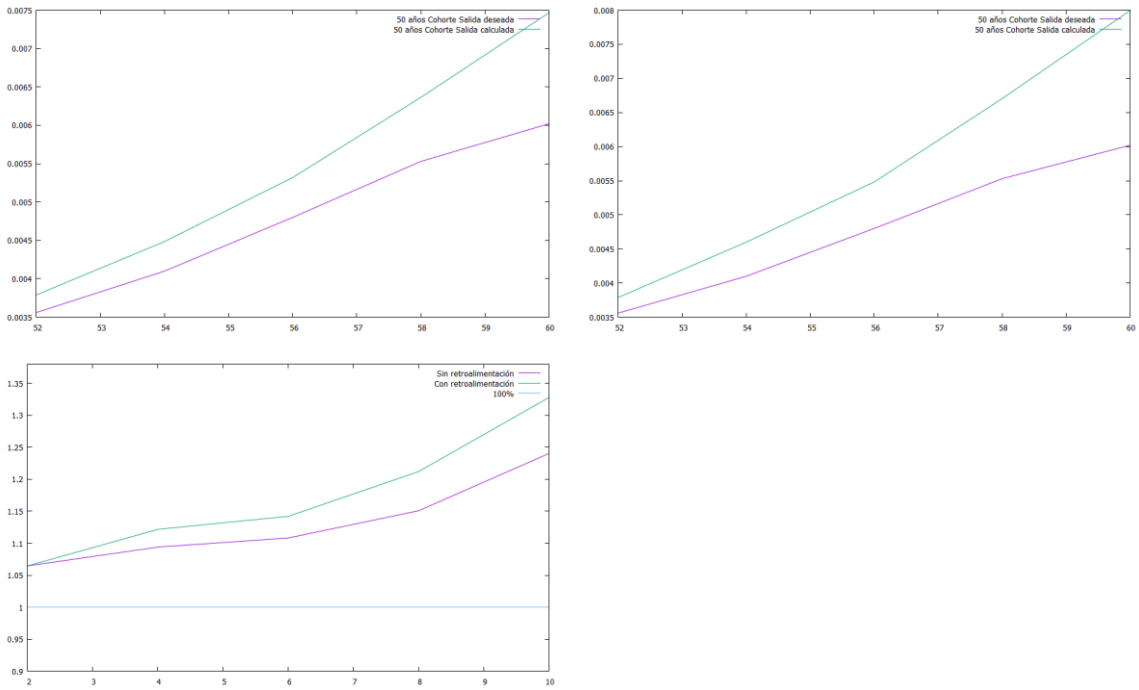
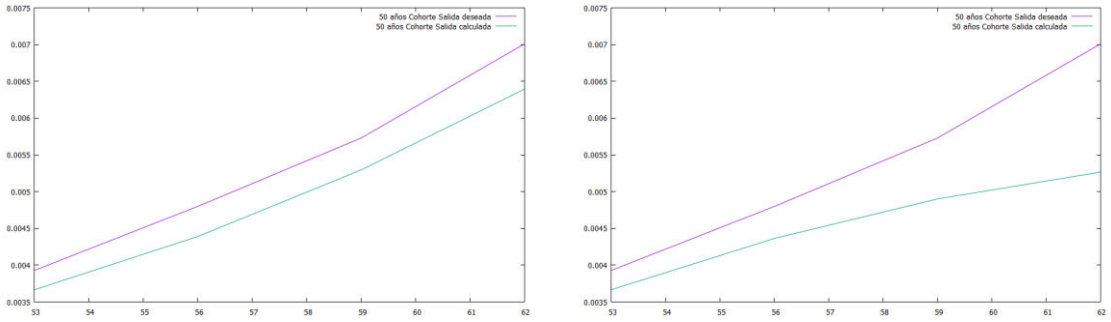


Figura 18: 50 años, Cohorte, pronóstico a dos años

El siguiente caso es el pronóstico a 3 años para el caso sin retroalimentar, aunque en este caso se tienen 4 pronósticos porque no se cuenta con datos suficientes en la tabla de mortalidad. En este caso se observa como la mortalidad es subestimada en todo el tramo de pronóstico y lo mismo sucede con la red retroalimentada con peores pronósticos.



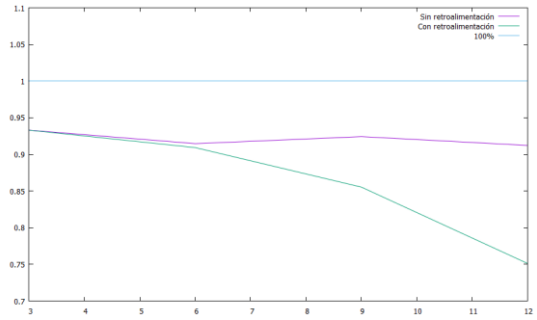


Figura 19: 50 años, Cohorte, pronóstico a tres años

En la siguiente imagen se realiza el experimento a 4 años vista, como en el anterior caso se cuenta con menos pronósticos, en este caso con 3 para las edades de 54, 58 y 62 años. Se comprueba que en este caso el error es bajo de hecho la edad intermedia de 58 años tiene más error que la de 62 años, y se observa como la red retroalimentada tiene más error que la no retroalimentada.

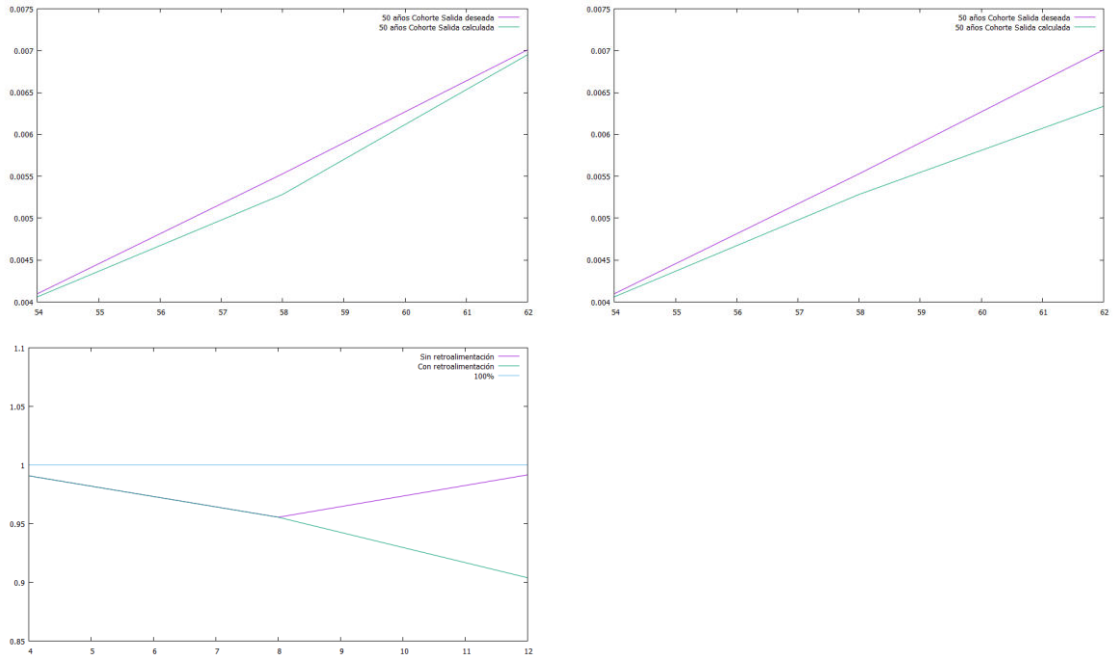


Figura 20: 50 años, Cohorte, pronóstico a cuatro años

Ahora llega el momento para los experimentos a 5 años, donde se tienen 2 pronósticos, el de 55 años y el de 60 años, donde se comprueba donde el error sube ligeramente a más edad de pronóstico, sobreestimando la probabilidad de muerte a un año. En el caso de la red retroalimentada, el error es algo mayor que en el caso de la red sin retroalimentar.

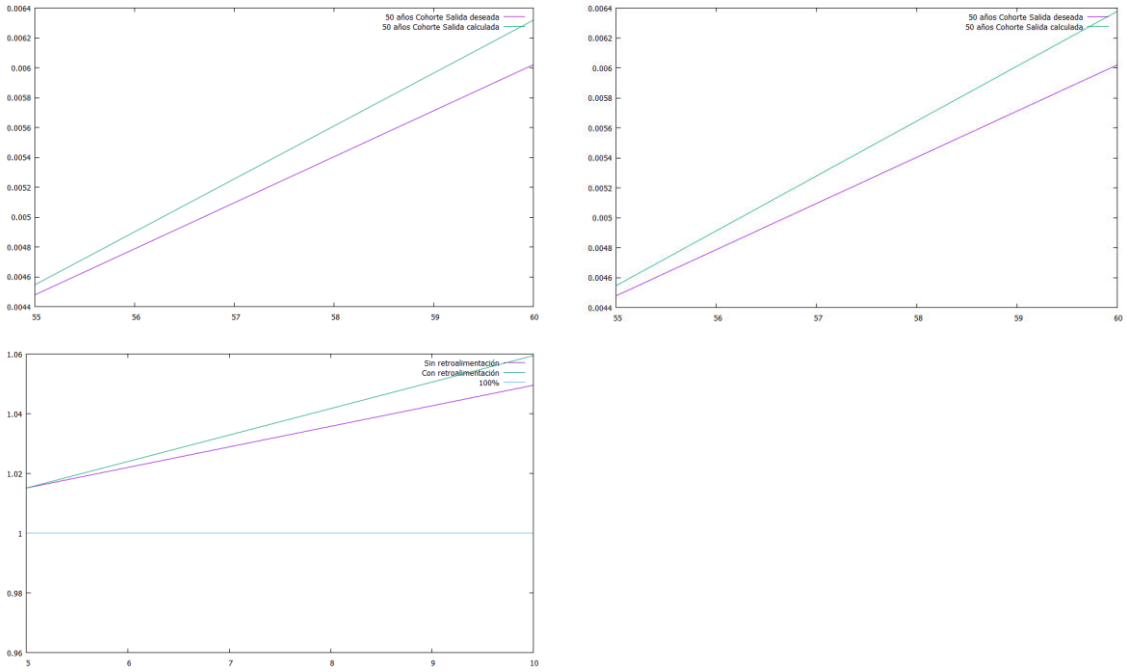


Figura 21: 50 años, Cohorte, pronóstico a cinco años

Se han realizado los mismos experimentos para el caso de 60, 70, 80 y 90 años, a continuación se ven las imágenes para el caso de red no retroalimentada (arriba a la izquierda), retroalimentada (arriba a la derecha), y se muestran la precisión en el pronóstico (abajo a la izquierda).

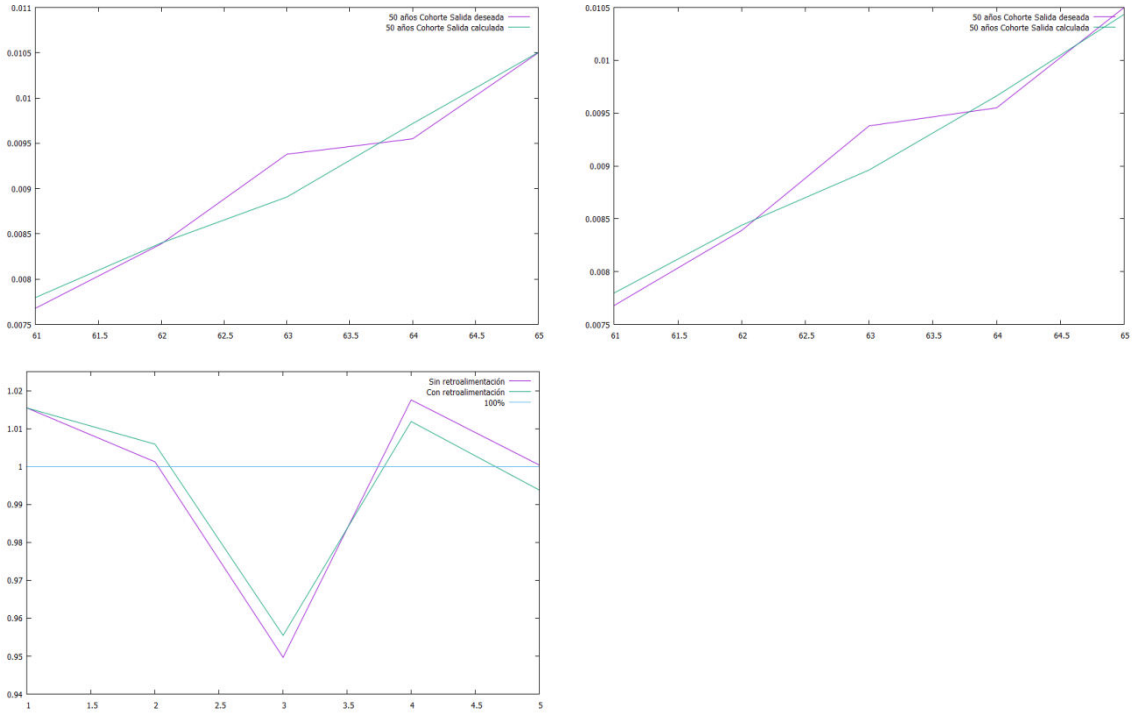


Figura 22: 60 años, Cohorte, pronóstico a un año

En este caso se ve que para pronósticos anuales, la red retroalimentada y la no retroalimentada van de la mano, siendo el error muy parecido en ambos casos.

Para el caso de pronósticos bienales, donde se pronosticará para 62, 64, 66, 68 y 70 años, se ve como el error va muy parejo en ambos casos, siendo algo mayor para el caso de red retroalimentada, mientras que una sobreestima la otra infraestima.

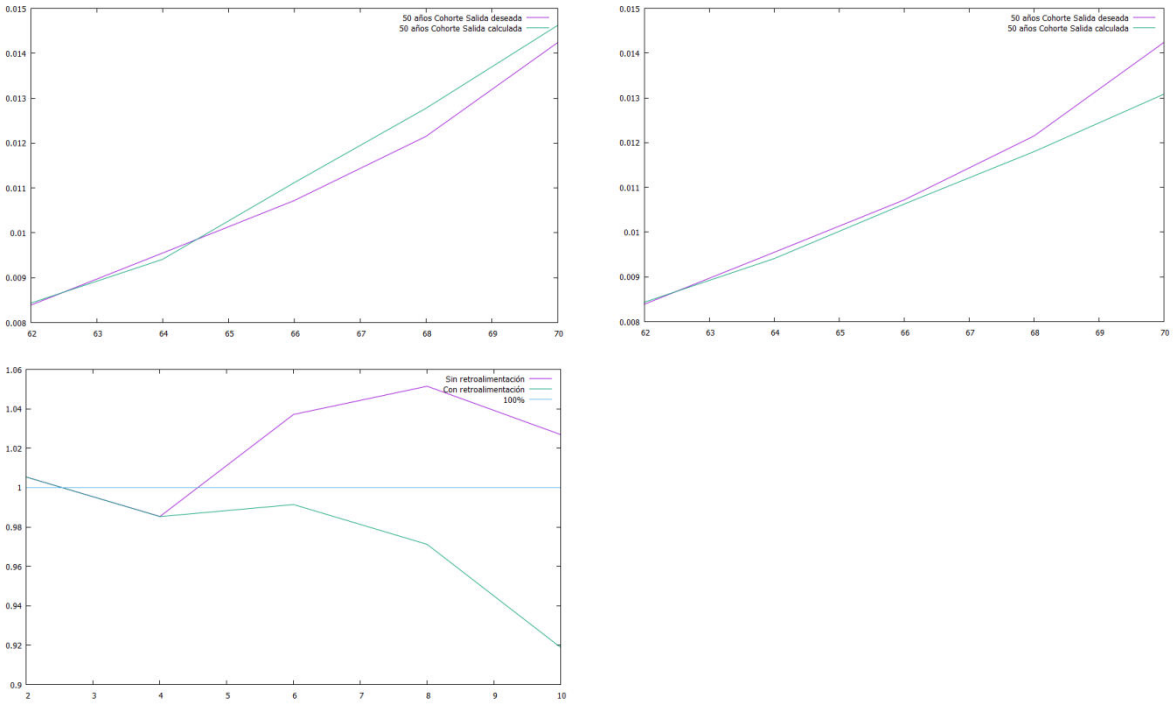
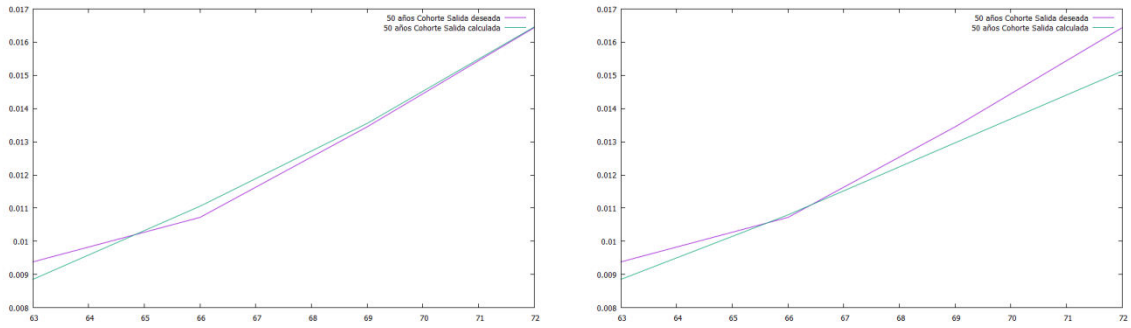


Figura 23: 60 años, Cohorte, pronóstico a dos años

En el caso de 3 años, los pronósticos son para 63, 66, 69 y 72 años, donde nuevamente el error cometido para el caso no retroalimentado es más bajo.



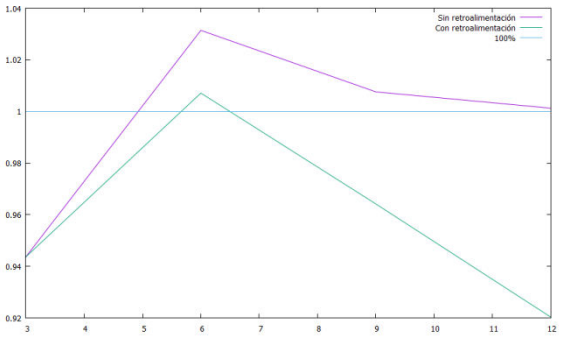


Figura 24: 60 años, Cohorte, pronóstico a tres años

Para el caso de pronósticos a 4 años los resultados vuelven a ser peores para la red retroalimentada.

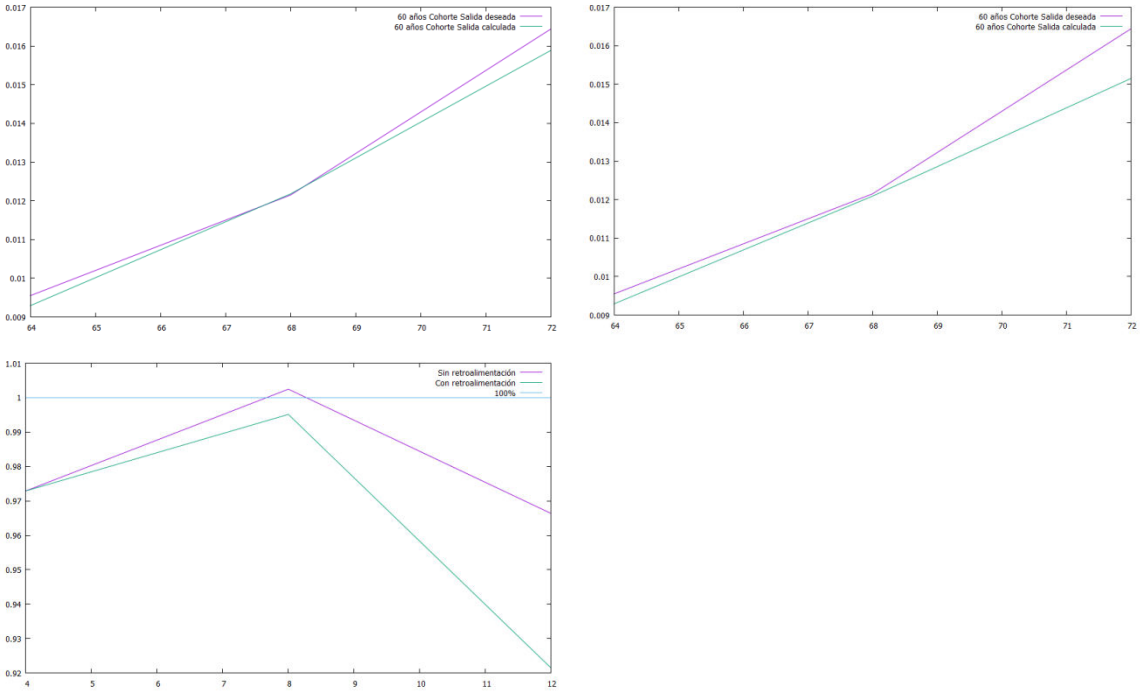


Figura 25: 60 años, Cohorte, pronóstico a cuatro años

Ahora se presentan los resultados para el caso de los pronósticos a 5 años, es decir para edades de 65 y 70 años. Se puede observar de nuevo como el error en el caso de red retroalimentada es algo peor, y que el error es menor a mayor periodo en ambos casos.

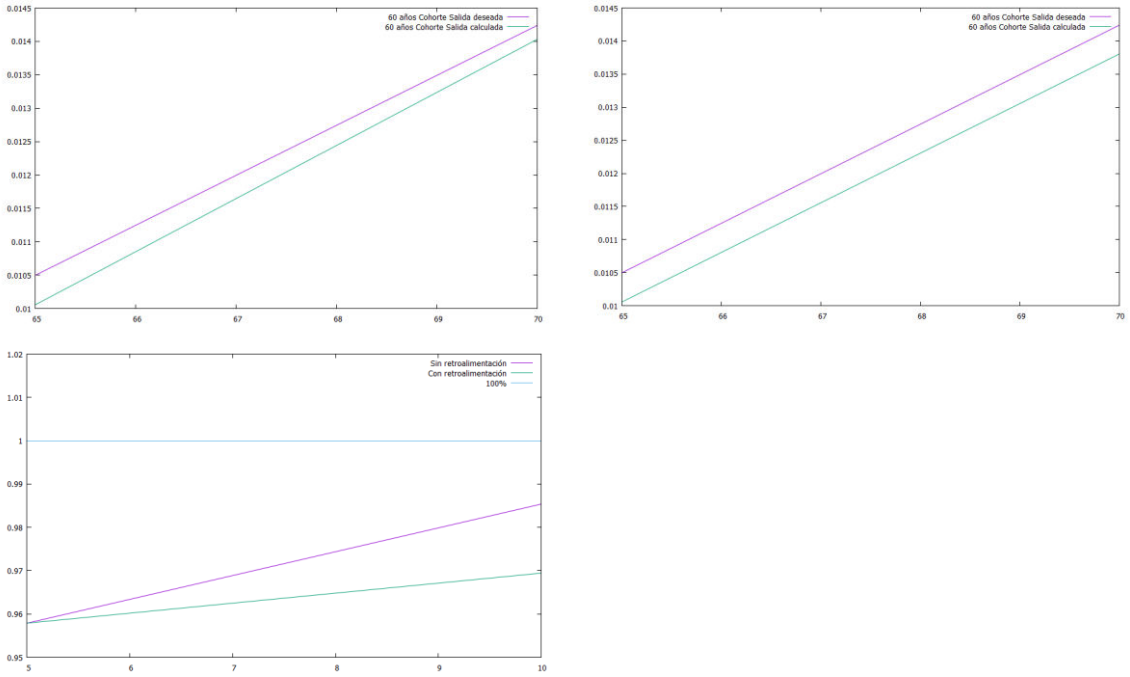


Figura 26: 60 años, Cohorte, pronóstico a cinco años

Ahora se muestran los experimentos realizados para la edad de 70 años, con pronósticos de un año. En este caso para ambos tipos los resultados son muy buenos.

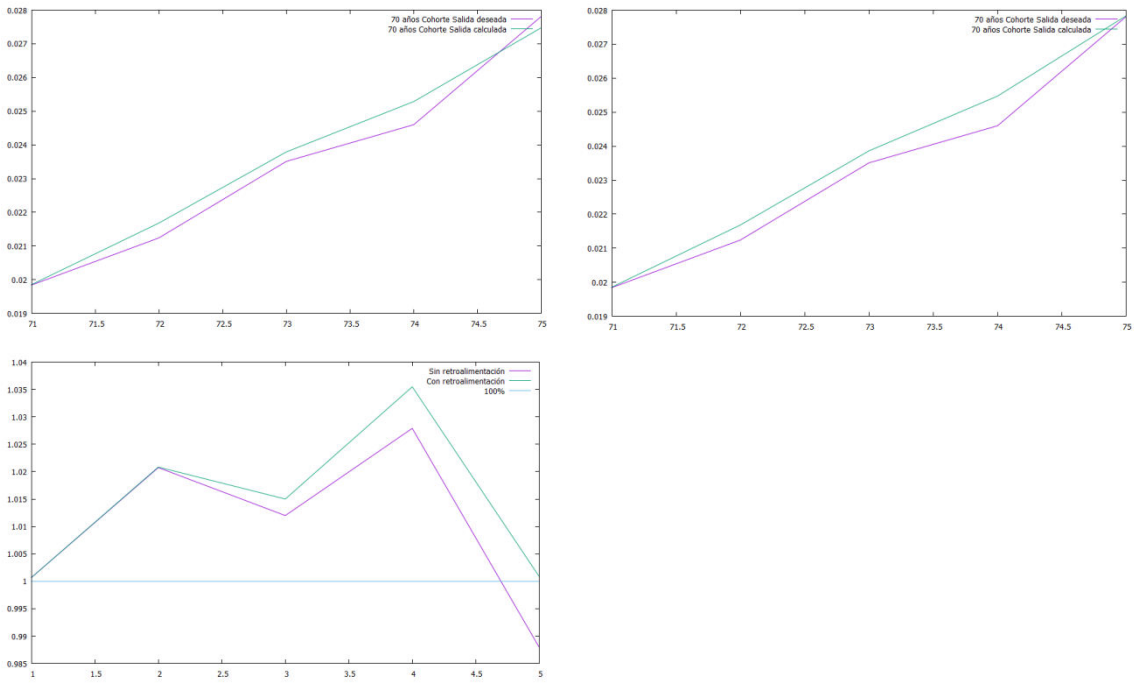


Figura 27: 70 años, Cohorte, pronóstico a un año

En el caso de pronosticar a dos años, los resultados son los que se ven más abajo, donde se puede apreciar que de nuevo el error de la red sin retroalimentación es mejor.

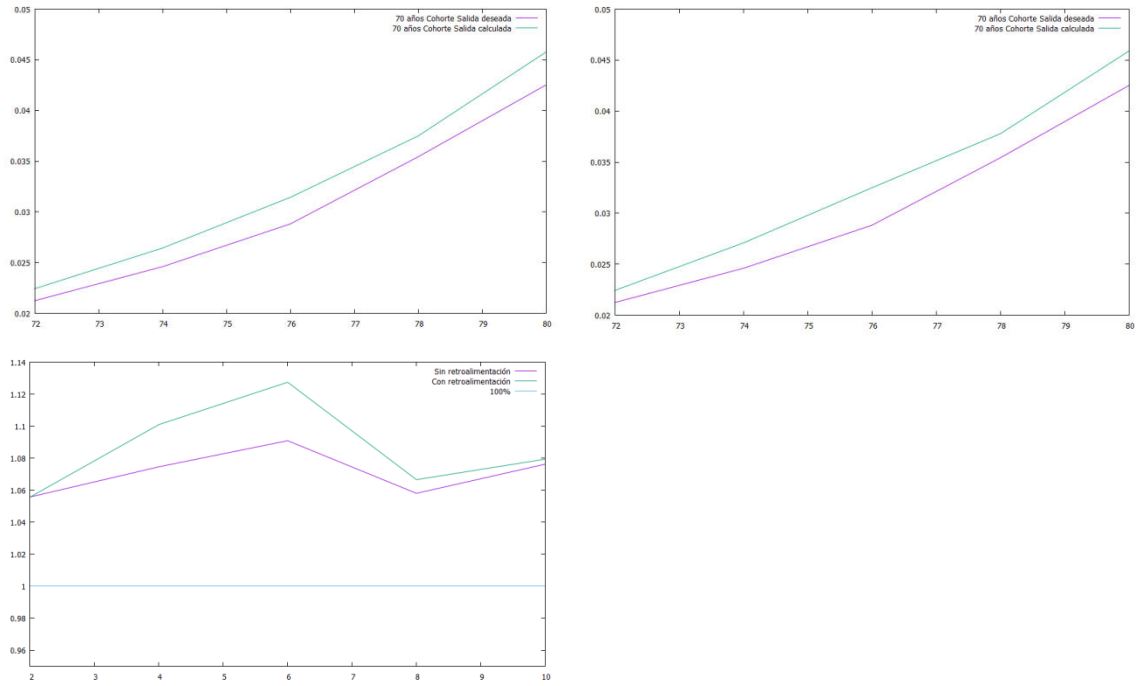
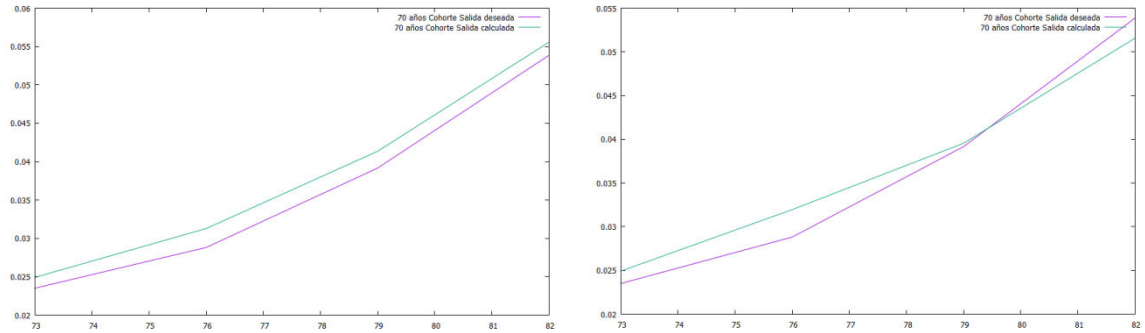


Figura 28: 70 años, Cohorte, pronóstico a dos años

Para el caso de pronóstico a tres años vemos que los resultados son muy buenos, y que por poco la red con retroalimentación tiene un error algo menor que la red sin retroalimentación.



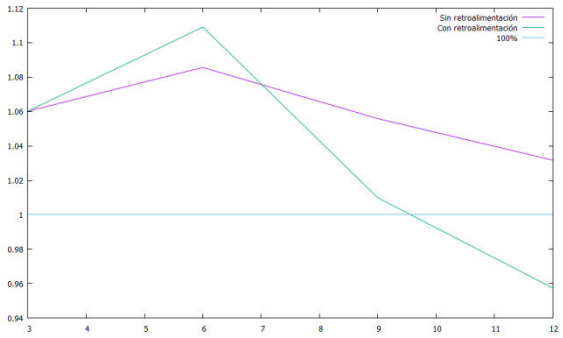


Figura 29: 70 años, Cohorte, pronóstico a tres años

En el caso de pronóstico a 4 años, donde de nuevo la red sin retroalimentación es mejor, siendo en este caso el pronóstico no tan bueno como en los casos anteriores.

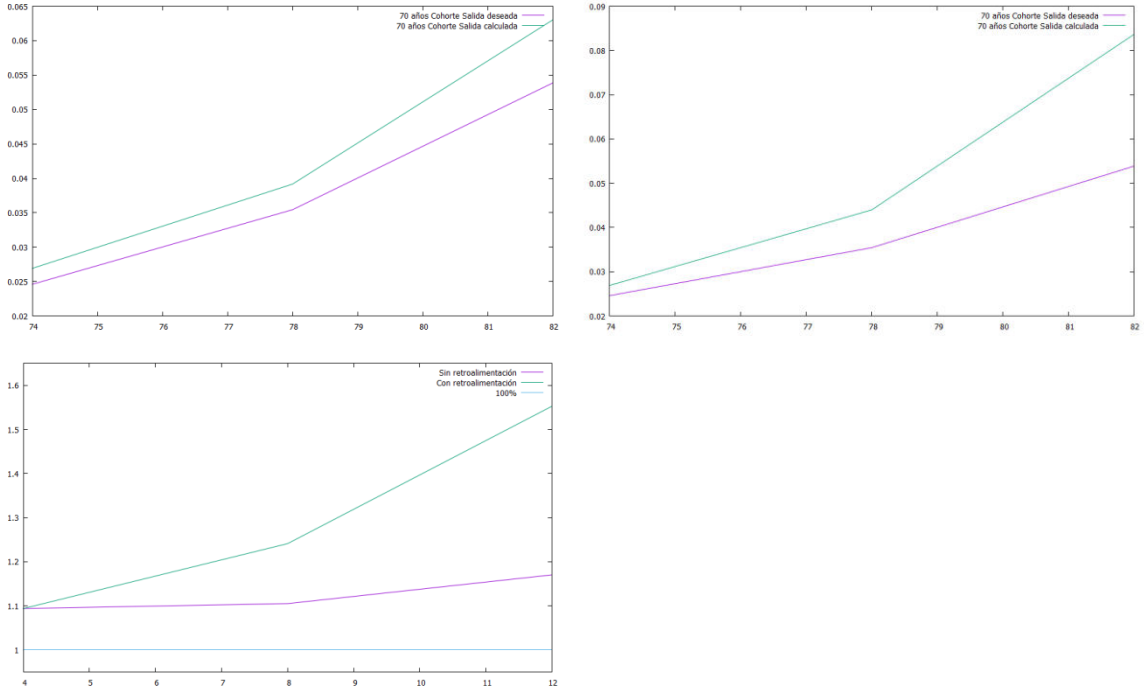


Figura 30: 70 años, Cohorte, pronóstico a cuatro años

Por último, en el caso de pronóstico a 5 años, de nuevo la red retroalimentada tiene los peores resultados, aún así aproxima bastante bien a esta edad.

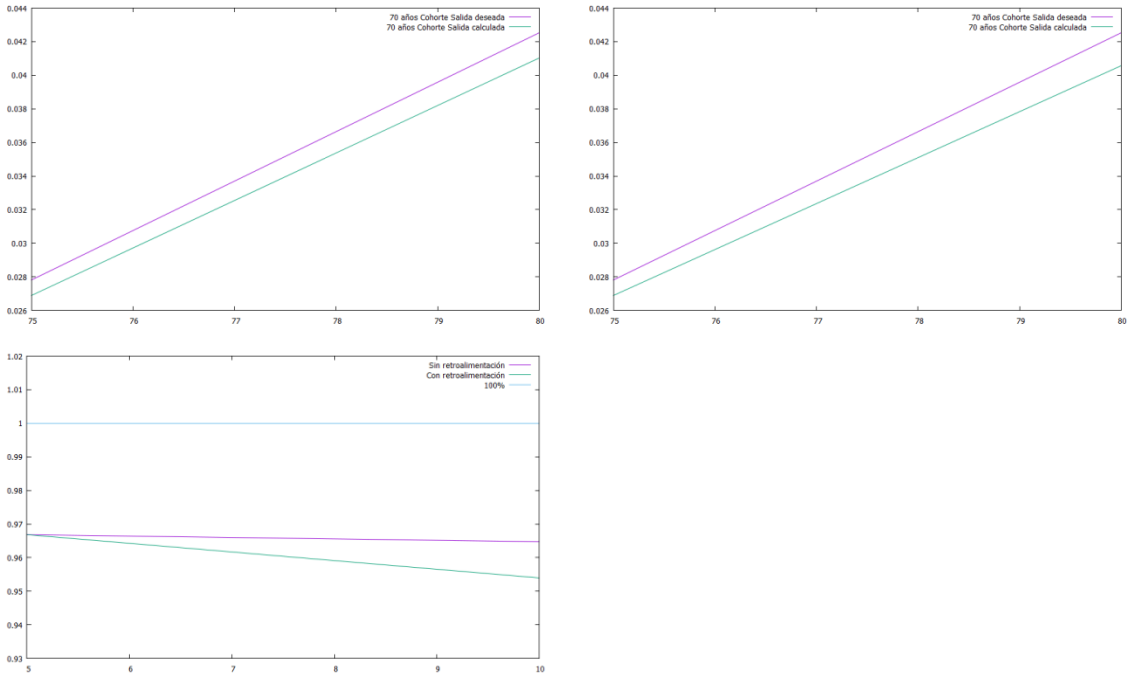


Figura 31: 70 años, Cohorte, pronóstico a 5 años

Ahora se muestran las gráficas para el caso de pronóstico para cohorte empezando en edad de 80 años, con pronóstico a un año. Como se comprueba en la gráfica, el error es muy alto en ambos casos, parece que para esta edad no es fácil aproximar con suficiente

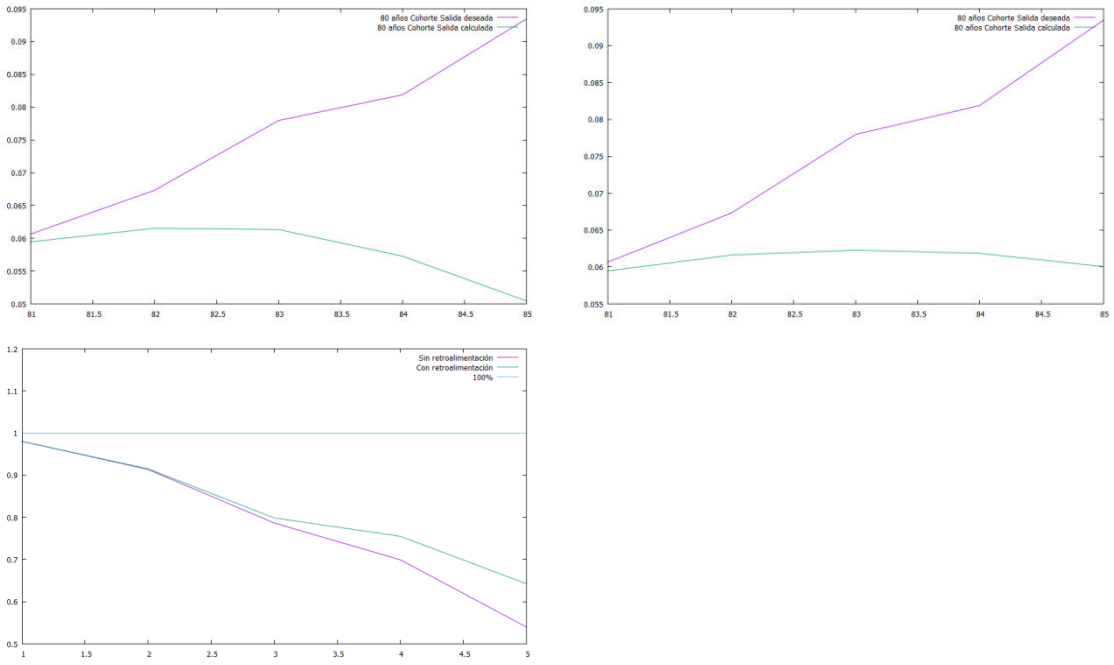


Figura 32: 80 años, Cohorte, pronóstico a un año

En el pronóstico a dos años, se ve como el error es muy grande, y va siendo mayor con cada pronóstico.

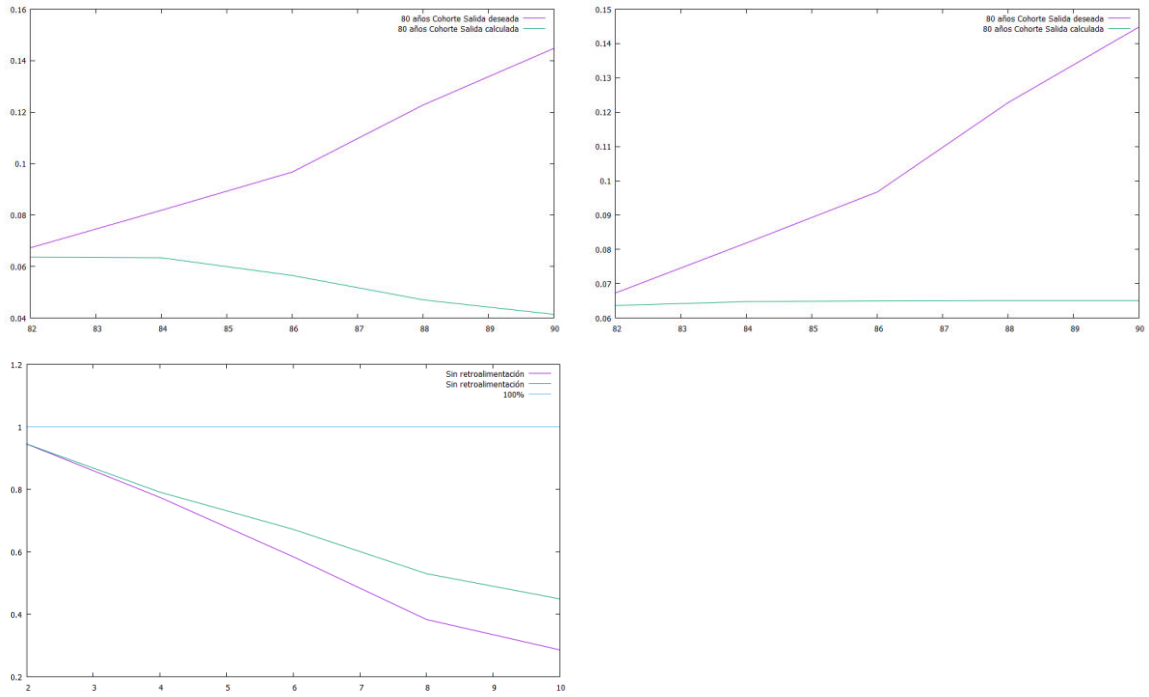
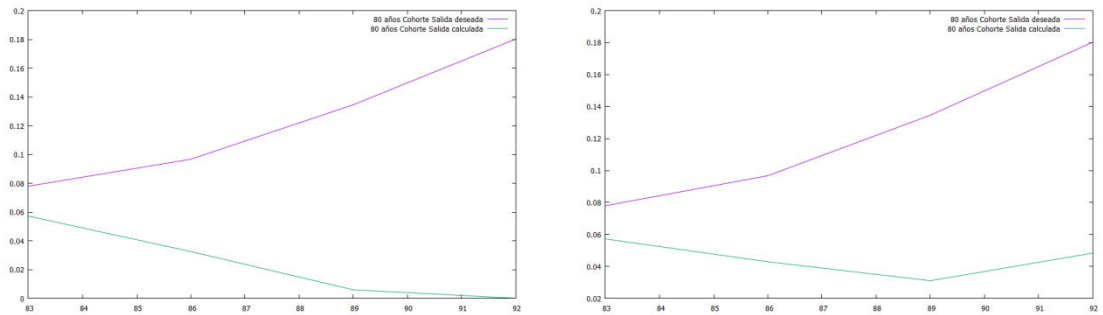


Figura 33: 80 años, Cohorte, pronóstico a dos años

Como se comprueba de nuevo, para el caso de pronóstico a 3 años, esta vez la red con retroalimentación tiene menos error que la no retroalimentada, en cualquier caso el error producido es muy alto.



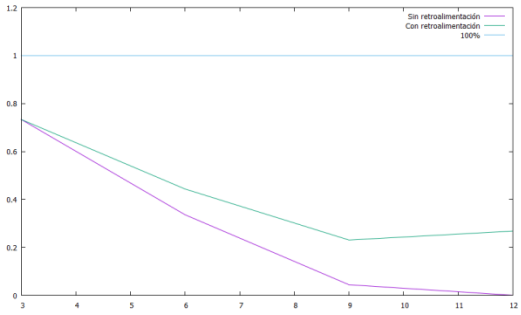


Figura 34: 80 años, Cohorte, pronóstico a tres años

En cuanto al pronóstico para 4 años, se vuelve a comprobar que el error es de nuevo muy alto. De nuevo error muy alto en ambos casos, y la red retroalimentada con mejor error que la no retroalimentada.

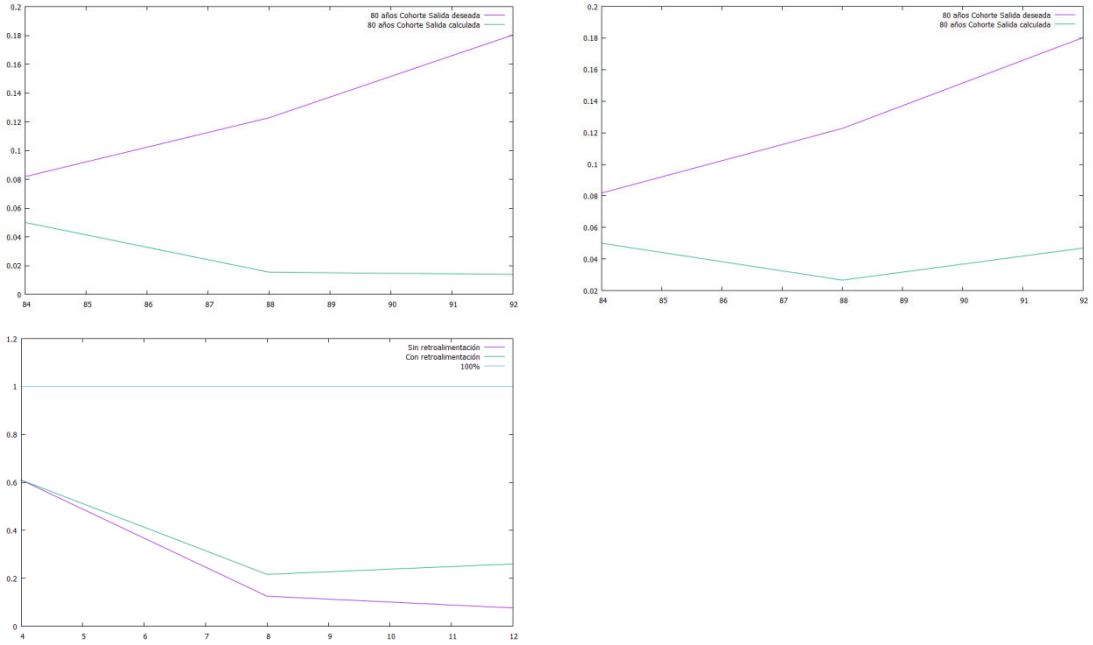


Figura 35: 80 años, Cohorte, pronóstico a cuatro años

Por último se pronostica a 5 años, dos pronósticos es decir las edades de 85 y 90 años, y de lo que se desprende en la gráfica podemos ver que el error sigue siendo muy alto, y que en este casi la red con retroalimentación sí tiene más error, aunque es bastante leve con relación a la red sin retroalimentación.

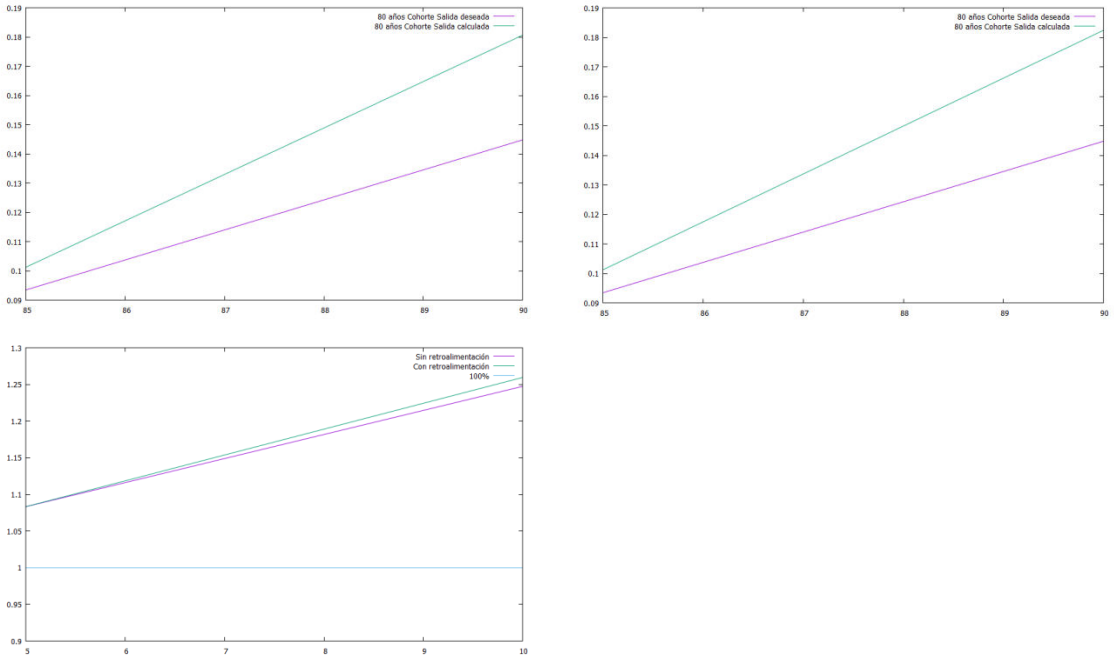


Figura 36: 80 años, Cohorte, pronóstico a cinco años

Por último, para la tabla de tipo cohorte, veremos la edad de 90 años, empezando por 5 pronósticos a un año, es decir los pronósticos para 91, 92… 95 años, vemos en la grafica de abajo que de nuevo los resultados contienen demasiado error.

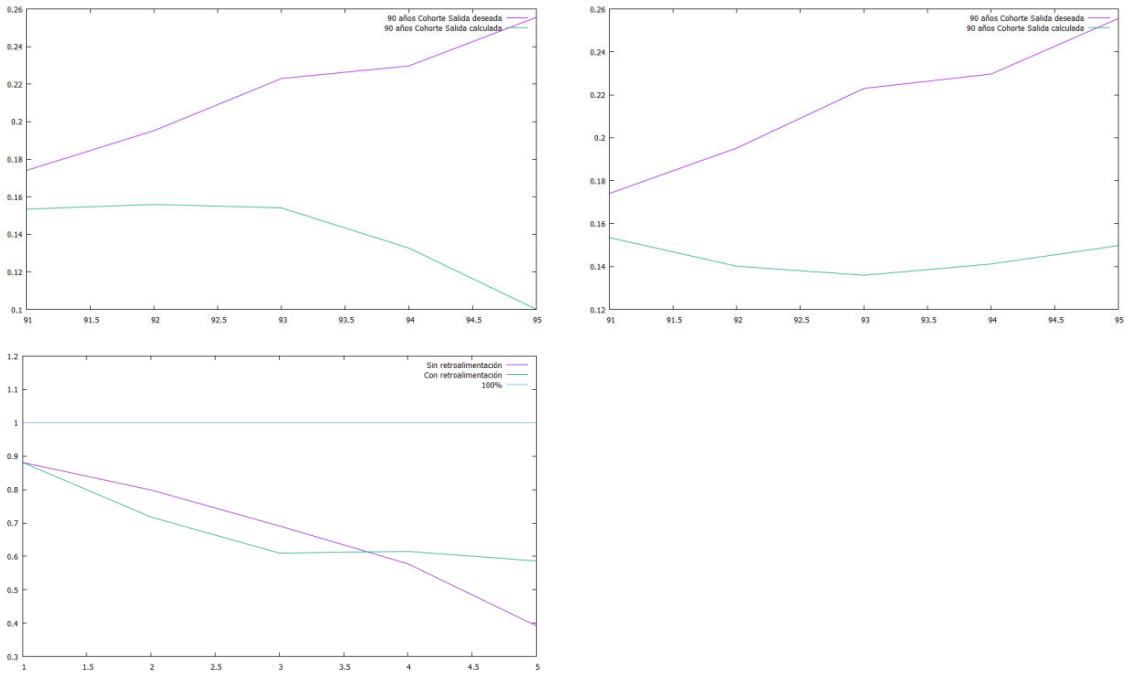


Figura 37: 90 años, Cohorte, pronóstico a un año

Ahora en el caso del pronóstico a dos años, viene a ocurrir más o menos lo mismo que con el caso anterior, error my alto. En este caso la red retroalimentada da menos error, pero aún así el error es grande en ambas.

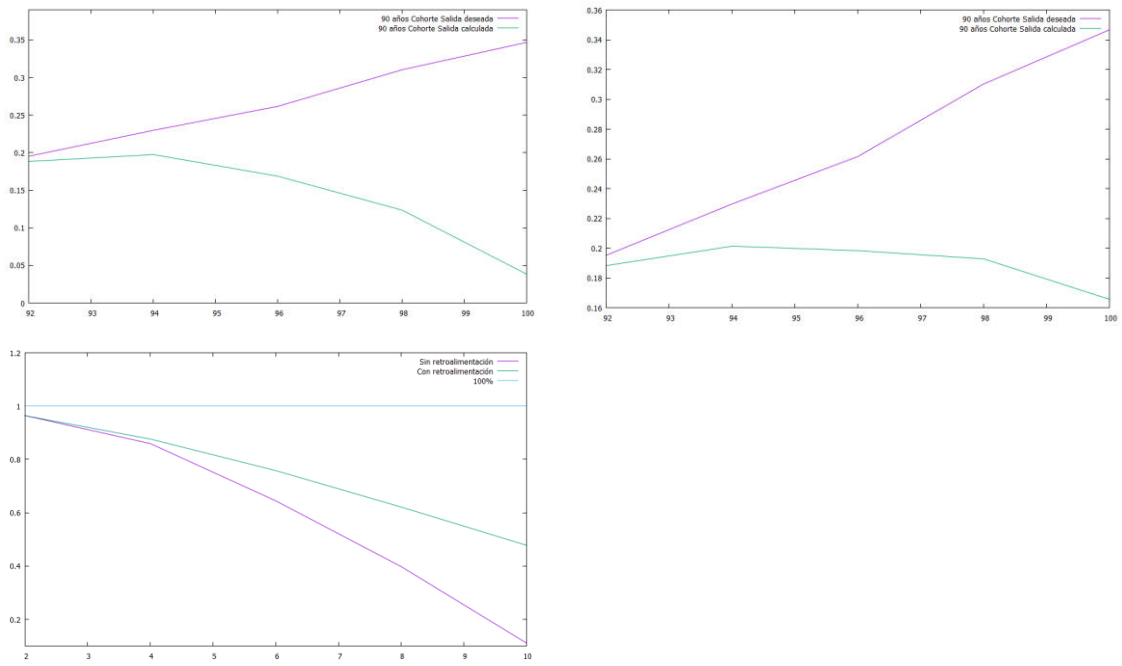
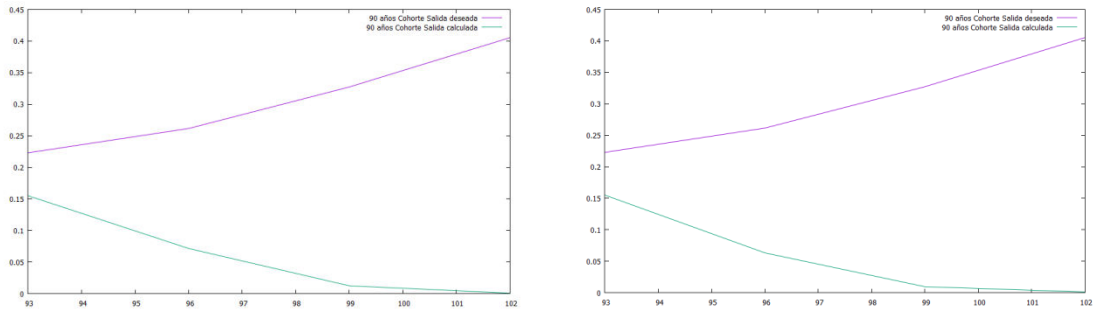


Figura 38: 90 años, Cohorte, pronóstico a dos años

En este caso probamos con pronósticos de 3 en 3 años, observando en las gráficas que tanto el error en el caso de la red con restroalimentación, como en el caso no retroalimentado, los errores son muy parecidos, pero muy altos, y se van haciendo más grandes con cada periodo pronosticado.



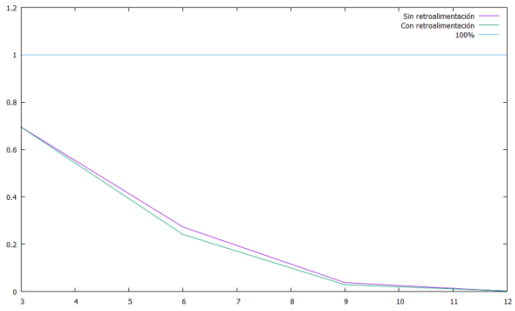


Figura 39: 90 años, Cohorte, pronóstico a tres años

En el pronóstico a 4 años podemos ver que la salida deseada y la calculada no tienen nada que ver, mientras una tiene una tendencia, la otra tiene la contraria.

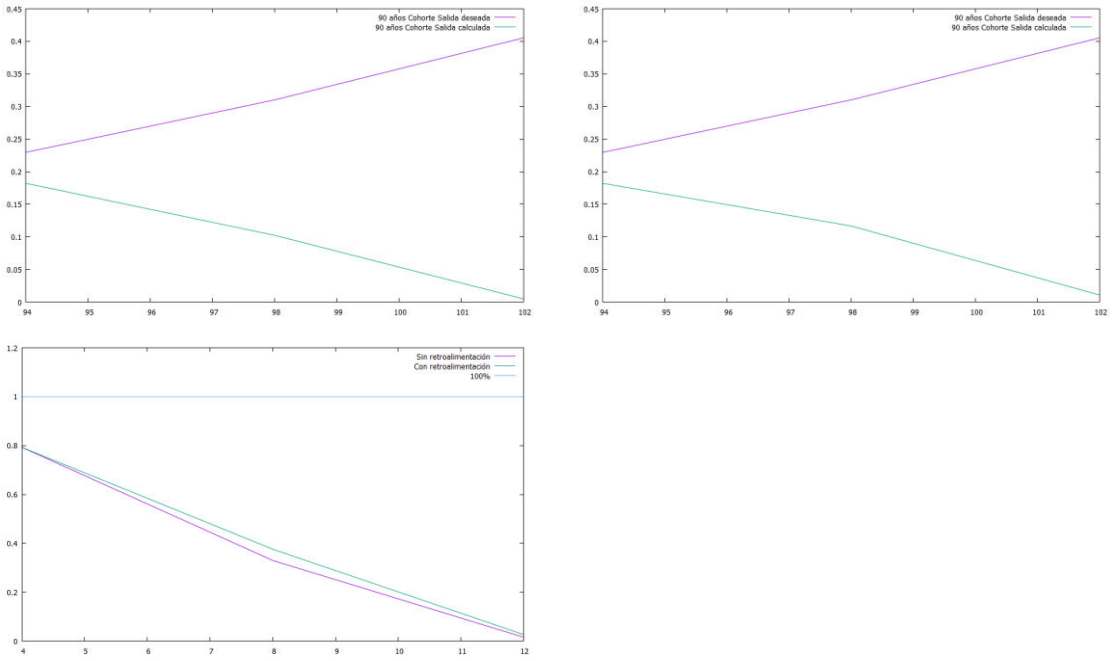


Figura 40: 90 años, Cohorte, pronóstico a cuatro años

Por último se prueba para pronosticar a 5 y 10 años, se puede ver que la tendencia es la correcta en el caso de la no retroalimentada, pero el error es muy grande como viene ocurriendo en el resto de los casos. Con la retroalimentada incluso la tendencia es errónea.

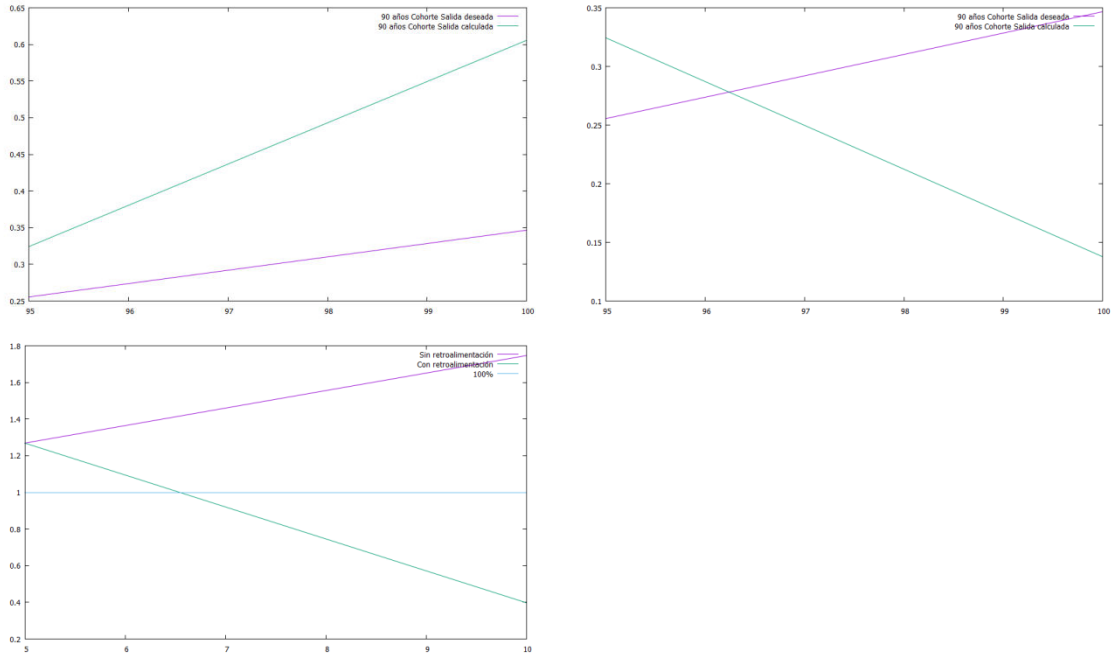
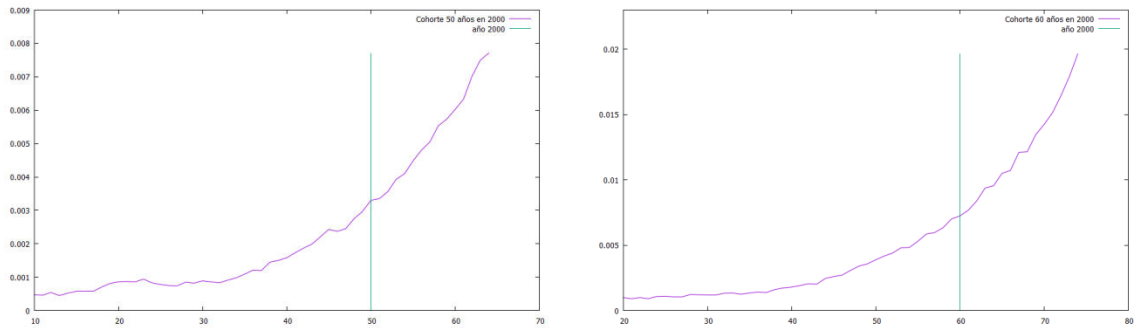


Figura 41: 90 años, Cohorte, pronóstico a cinco años

Esta fue una batería de pruebas lanzada, y que terminó varios días después, y como se ha podido verificar, los datos para edades avanzadas no eran buenos. Tenía que existir alguna explicación para esto, quizá es que los valores a esas edades tuvieran muchos altibajos y por eso la red no fuera capaz de generalizar, o que las variaciones entre cada dato fueran muy bruscos.

También pudiera ser que el número de neuronas en este caso no fuera suficiente con 10 en la capa oculta, desde luego los cambios que se podían realizar en los parámetros eran muchos, pero para eso lo mejor era recoger los datos de la tabla de cohorte para las edades consideradas y ver su variación.



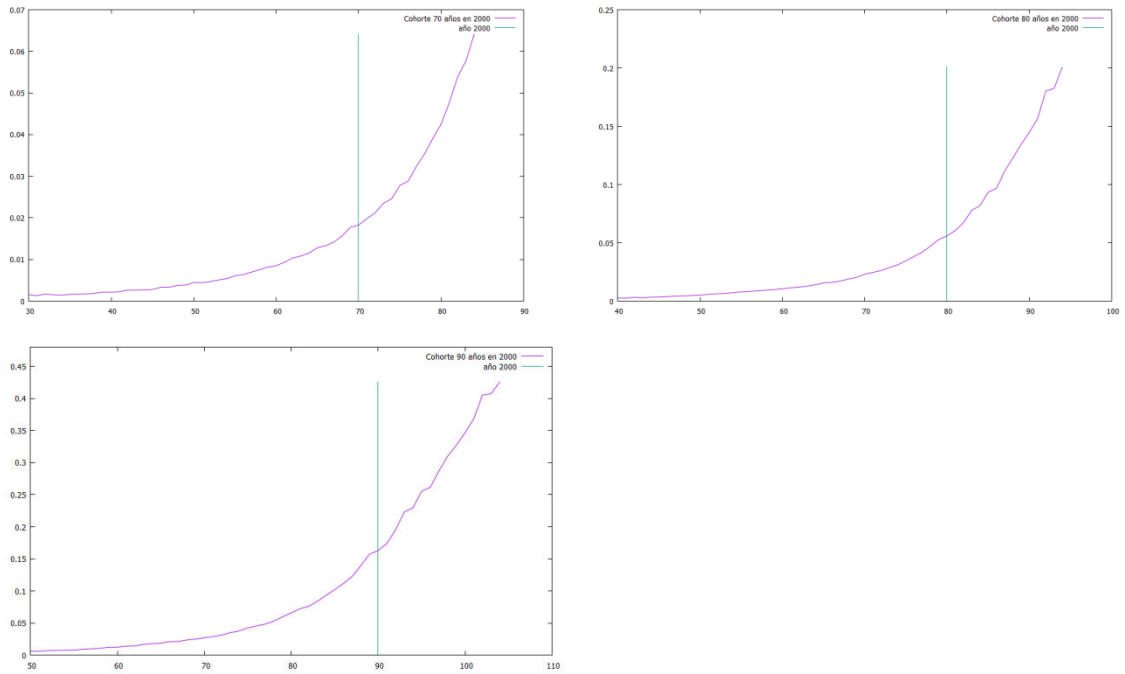


Figura 42: Mortalidad Cohorte para 50, 60, 70, 80 y 90 años en el 2000

Ciertamente no se observa en principio, nada que pueda inducir a pensar que existe algún problema en las graficas. Es cierto que la pendiente se hace más abrupta a partir del punto de corte, que es el punto hasta donde se entreno la red, sobre todo en el caso de edades altas, pues la mortalidad aumenta aceleradamente en la zona de pronóstico.

Lo que se determinó es que quizá hubiera que coger un menor número de elementos de validación, porque el cambio lo introdujeran realmente un menor número de patrones, pero esto no dio resultados. Lo siguiente a probar es recortar el número de patrones de entrenamiento, persiguiendo lo mismo, que realmente se tuvieran en cuenta los patrones con una mayor semejanza entre ellos, al fin y al cabo, los valores a pronosticar son más parecidos a las edades un poco anteriores, que a edades de 40 años menos, así que se decidió iniciar el entrenamiento en el año 1980 en vez del 1960 y con un elemento de validación, que es la última edad pronosticada.

En este caso se probó para 80 años con pronóstico a un año, cinco pronósticos, y se comprobó como la precisión aumentó ligeramente. El siguiente paso era aumentar el número de neuronas en la capa oculta hasta un máximo de 15 y los patrones de validación a 4. Aquí ya la precisión observada ya si mejoró radicalmente, demostrando

que las redes de neuronas en general son muy sensibles a las parametrizaciones, y este tipo de red en concreto tiene muchos grados de libertad.

Una vez conseguido esto se volvió a entrenar de forma automática una serie de redes de neuronas, y seleccionando como mejor a la que menor error de validación tenía. De nuevo los valores fueron mucho mejores que en el caso de entrenar desde el año 1960. Se probó como incidía la selección del número de patrones de validación, y se vio que lo mejor se encontraba entre 4 y 5 elementos de validación, mayor o menor número a estos no hacían sino disminuir la precisión del pronóstico. Lo mejor es que los resultados los ofrecía de forma consistente, siempre moviéndose entre un 3%-4% por arriba o por debajo, es decir, que se lanzaron 5 pruebas para el caso de 4 elementos de validación, y otras cinco con 5 elementos de validación, y las redes devueltas como mejores resultados, arrojaban consistentemente estos valores para el error no ya de la red sin retroalimentar, sino para la red retroalimentada.

En este punto se puede comprobar como una parametrización en concreto puede tener tanta repercusión. Y no solo los elementos más comunes, como pueda ser la razón de aprendizaje, el tipo de entrenamiento, la función de activación o el número de neuronas, sino el número de elementos de validación tuvo una repercusión muy importante. Por fortuna se escogieron 4 elementos de validación desde el principio, pero habría que estudiar la repercusión a distintas edades.

La precisión alcanzada para la edad de 80 años, pronosticando las edades de 81, 82, 83, 84, y 85 años son los que aparecen en la gráfica de abajo, donde a la izquierda aparece la precisión de la red sin retroalimentar frente a la retroalimentada, donde se comprueba una precisión muy buena, para 4 elementos de validación. A la derecha lo mismo pero para 5 elementos de validación, donde se observa que en ambos casos como el error fluctúa entre el 3% o 4% por arriba o por debajo del valor real.

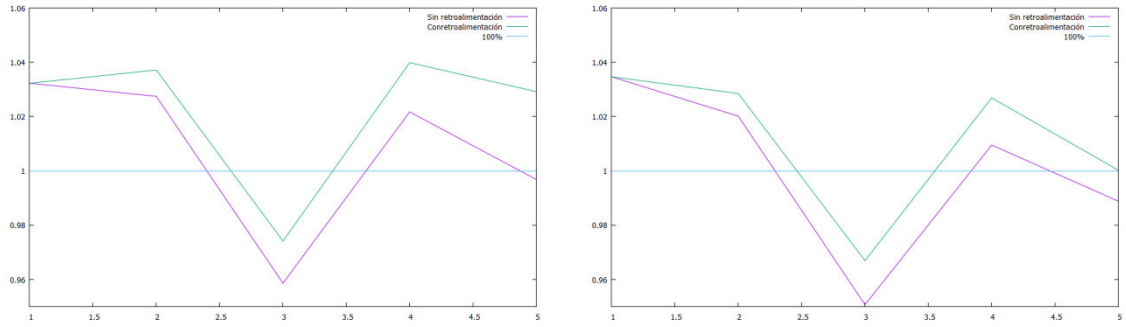
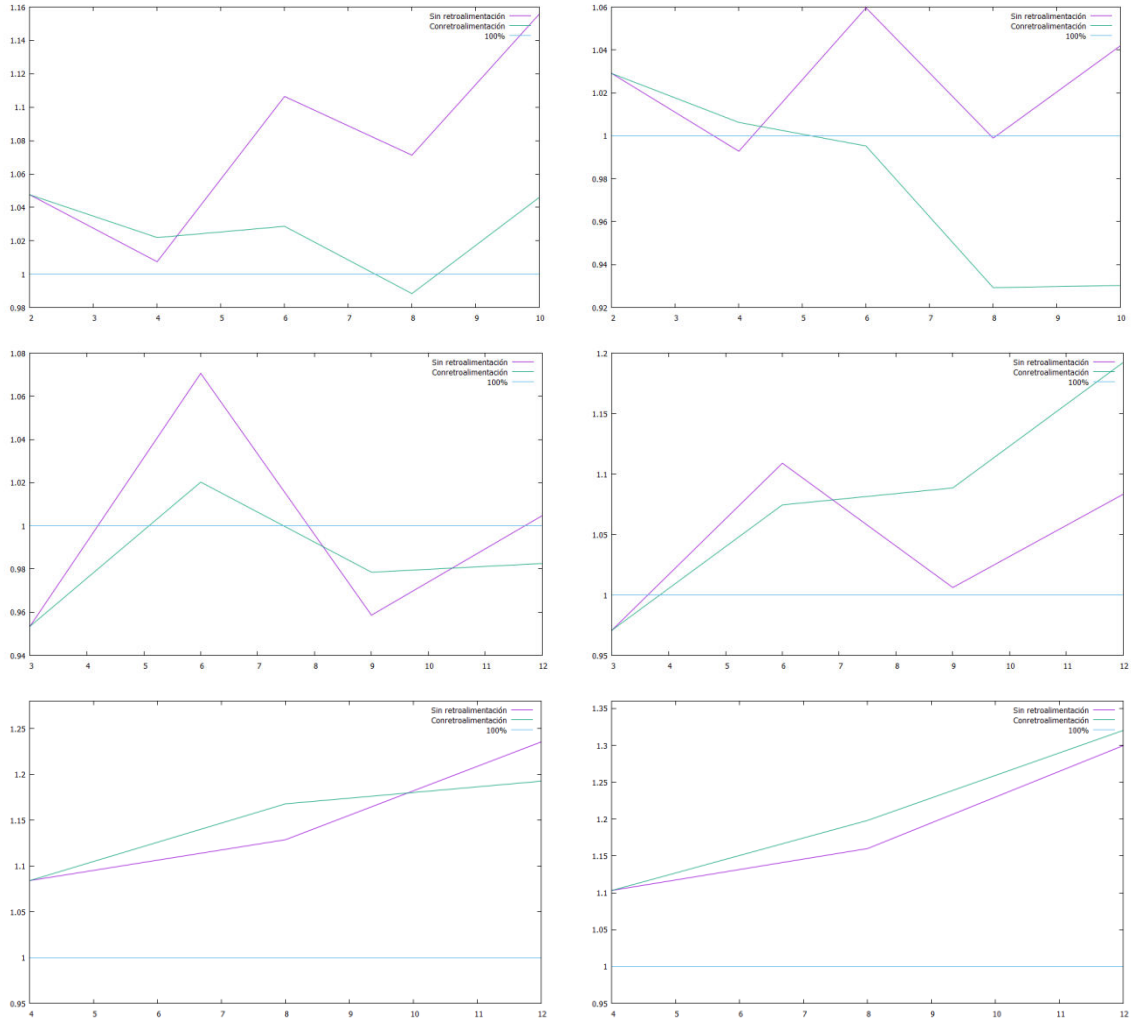


Figura 43: Precisión para 80 años Cohorte 5 pronósticos a un año

A continuación se muestran los resultados para 4 y 5 elementos de validación de izquierda a derecha, y de arriba abajo para 2, 3, 4 y 5 años de pronóstico.



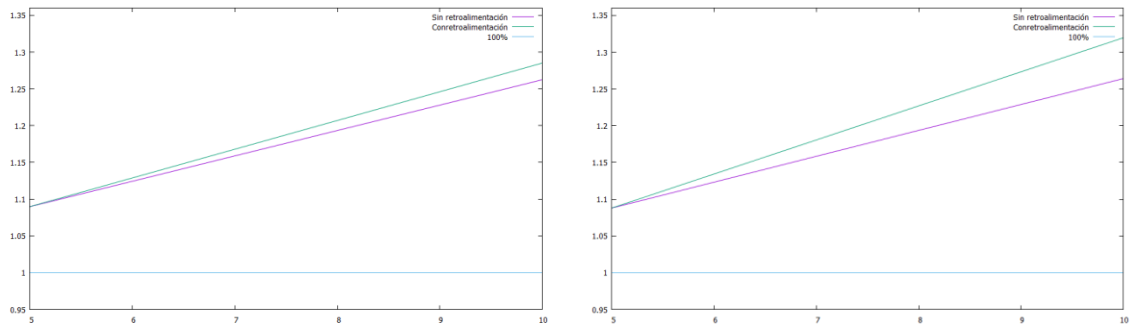


Figura 44: Precisión para 80 años Cohorte 5 pronósticos a 2, 3, 4 y 5 años

Como se puede observar en la figura 44 los errores aumentan con el año de pronóstico pero se observa algo curioso, y es que por ejemplo si se observa lo que devolvía la red que pronosticaba a un año en el quinto año, el error es menor que el primer pronóstico de la red que pronostica a cinco años, es decir que para la misma edad pronosticada, la red de pronóstico a un año lo hace mejor que la de pronóstico a cinco años.

Introduciendo menos tramo de entrenamiento, esta vez con patrones que van desde 1985 al 2000 se ha conseguido el siguiente error, para validación con 4 y 5 elementos, lo cual es un error muy bueno a 5 y 10 años, el problema es que de uno a otro lanzamiento hay cierta disparidad en los resultados.

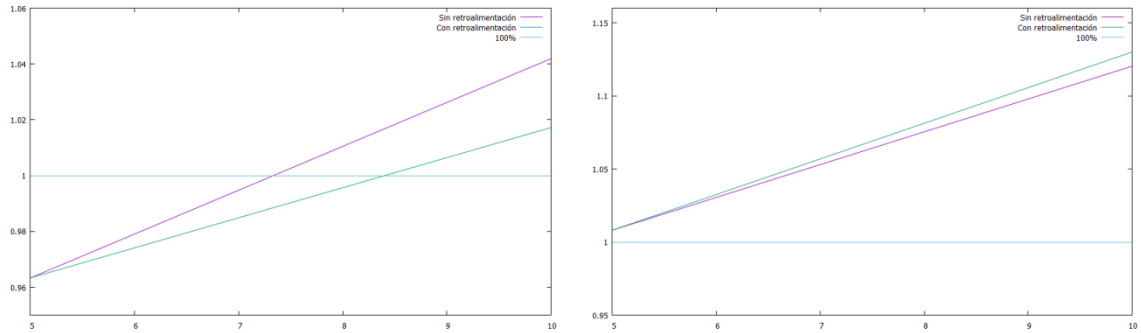


Figura 45: Precisión para 80 años Cohorte 5 pronósticos a 5 años

En cuanto a los pronósticos para la edad de 90 años sí mejora la precisión de pronóstico pero no al nivel de lo ocurrido con los 80 años. En las siguientes gráficas se pueden ver de izquierda a derecha y de arriba hacia abajo, la precisión en la previsión a un año, a dos… y a cinco años. Donde el periodo de entrenamiento seleccionado es desde 1980 al 2000, sorprende la precisión de pronóstico a cinco años, la cual es muy buena.

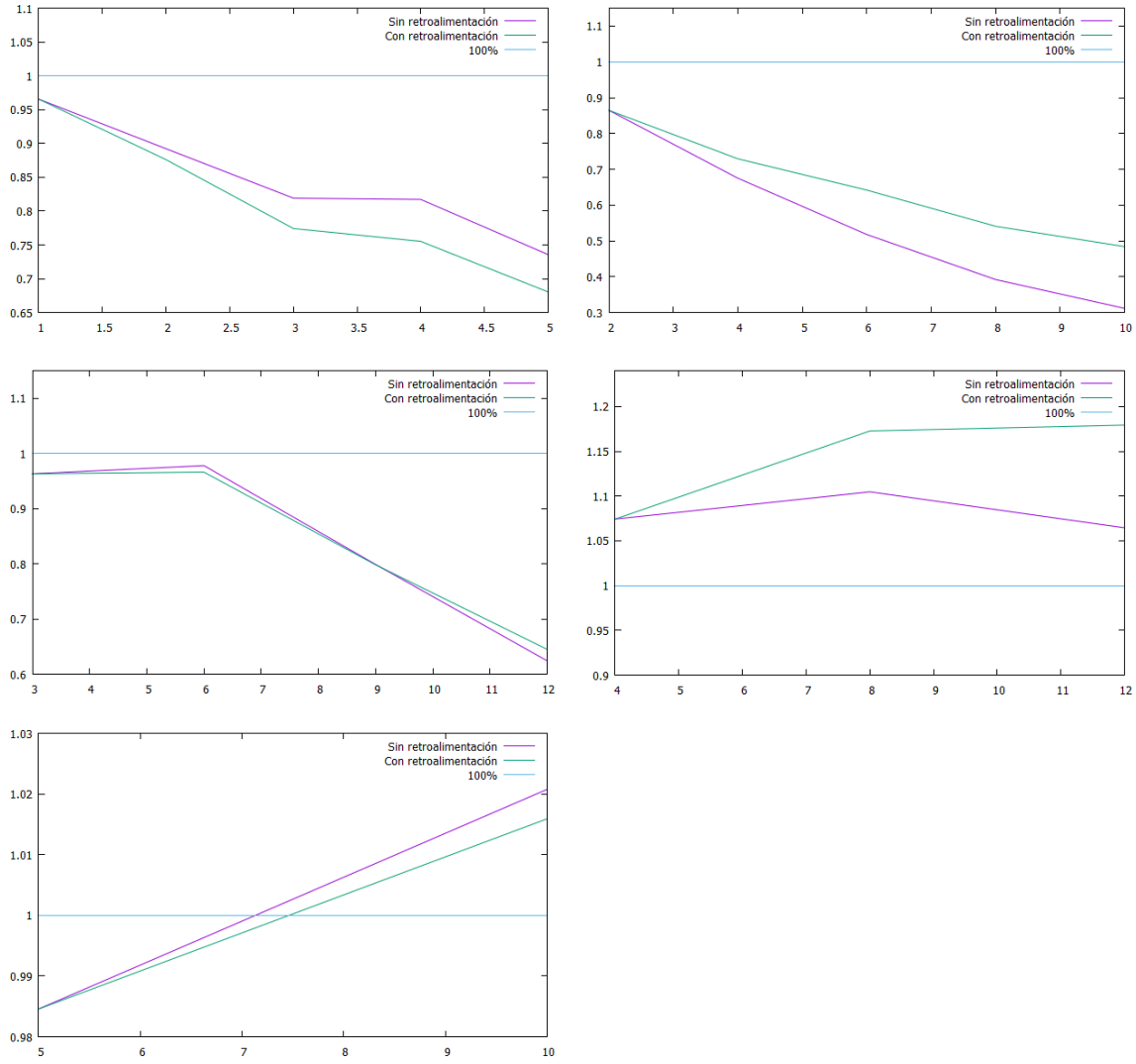


Figura 46: Precisión para 90 años Cohorte pronóstico [1-5] años

El problema es que no se produce una regularidad en las salidas, es decir que si se vuelve a lanzar el experimento puede dar mejor o peor valores de pronóstico, en torno a un 10% por arriba o por abajo, cosa que con la edad de 80 años no pasaba. En este caso se ha probado con la función de activación multicuadrática inversa, que parece mejorar algo los resultados en relación a la gaussiana.

Ahora se mostrarán las precisiones en el pronóstico para cada una de las edades consideradas, que son las de 50, 60, 70, 80 y 90 años de tipo periodo.

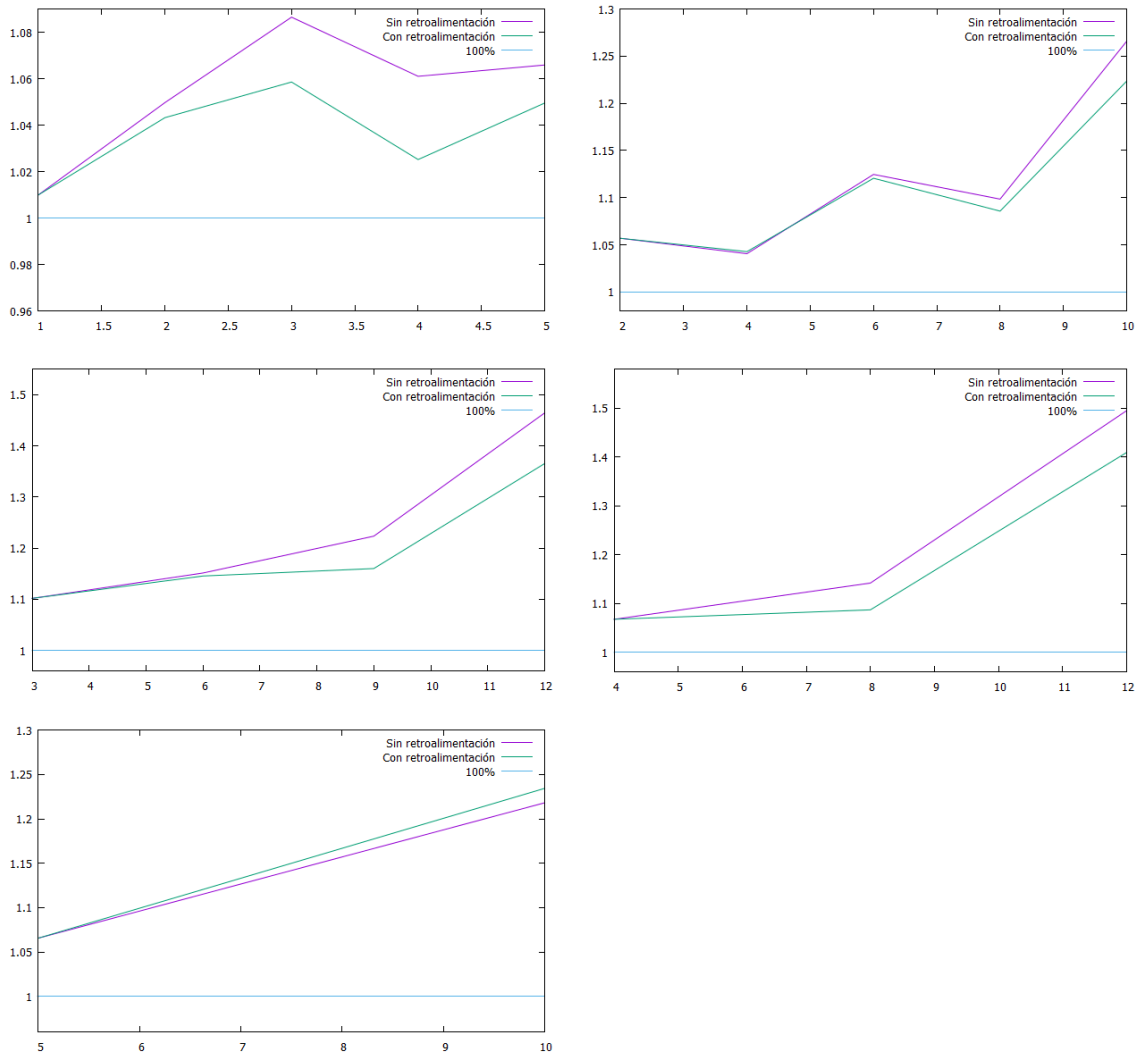


Figura 47: Precisión para 50 años Periodo pronóstico [1-5] años

Como se puede observar, el pronóstico a un año durante cinco años es muy preciso, no superando el 5% de error, y siendo mejor el pronóstico mediante retroalimentación que sin ella. Paulatinamente el error va siendo peor hasta llegar a un 40% en el caso del pronóstico a 4 años cuando se pronostican los 12 años.

En cuanto a los pronósticos para los 60 años, las gráficas con la precisión son las que se muestran abajo, donde se puede ver como los resultados son mejores que para 50 años, como ocurría al tratar la cohorte.

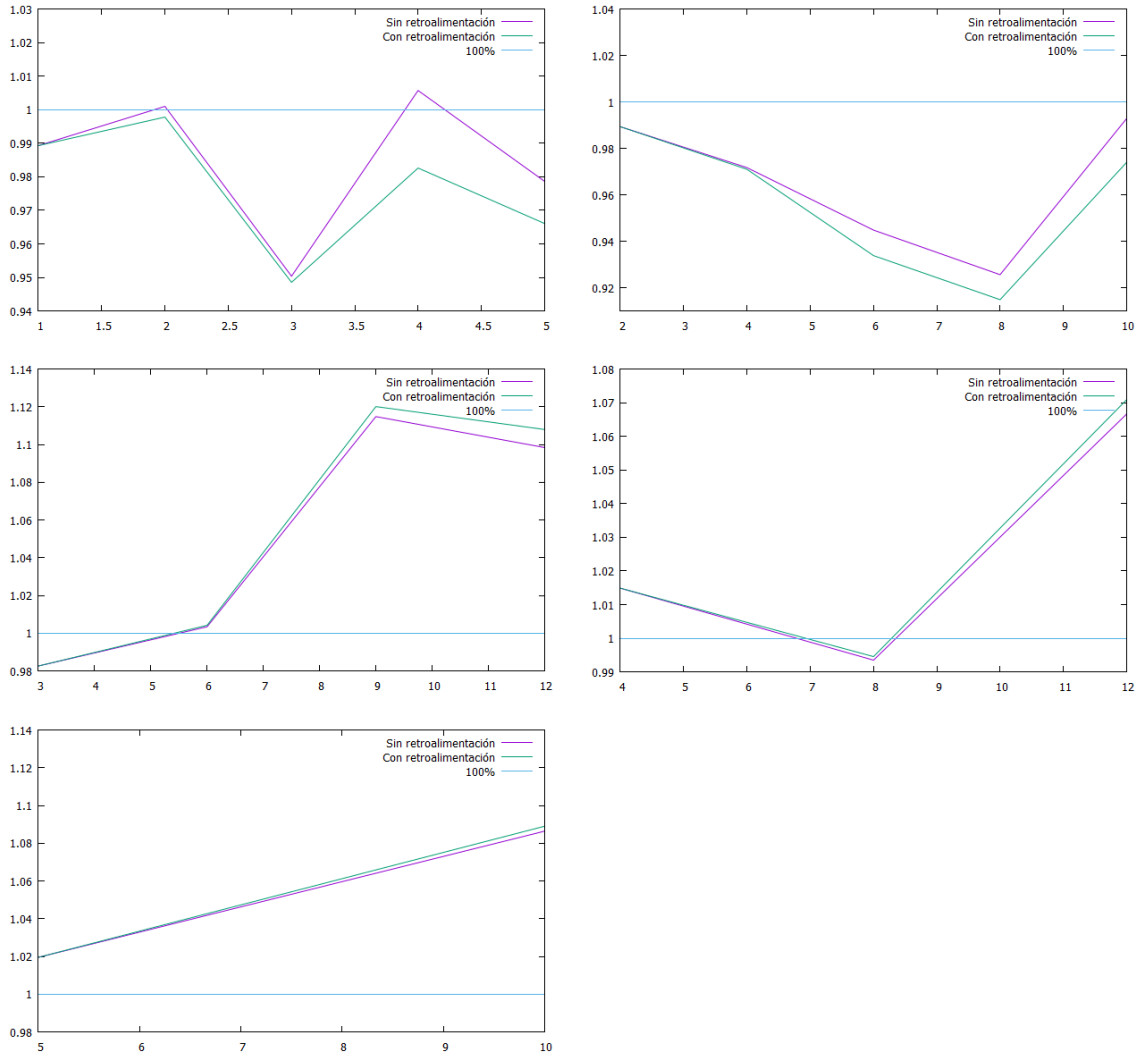
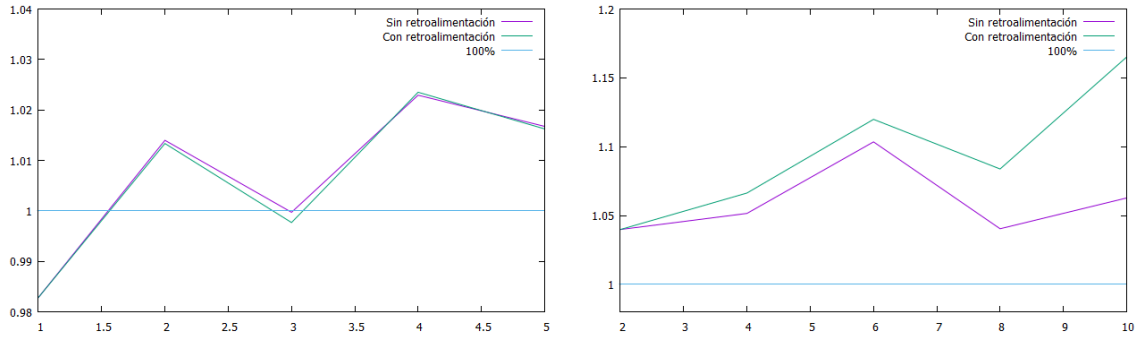


Figura 48: Precisión para 60 años Periodo pronóstico [1-5] años

A continuación se muestra las gráficas de los pronósticos para 70 años, donde se puede ver que extrañamente el pronóstico a 2 y 3 años son peores que el resto.



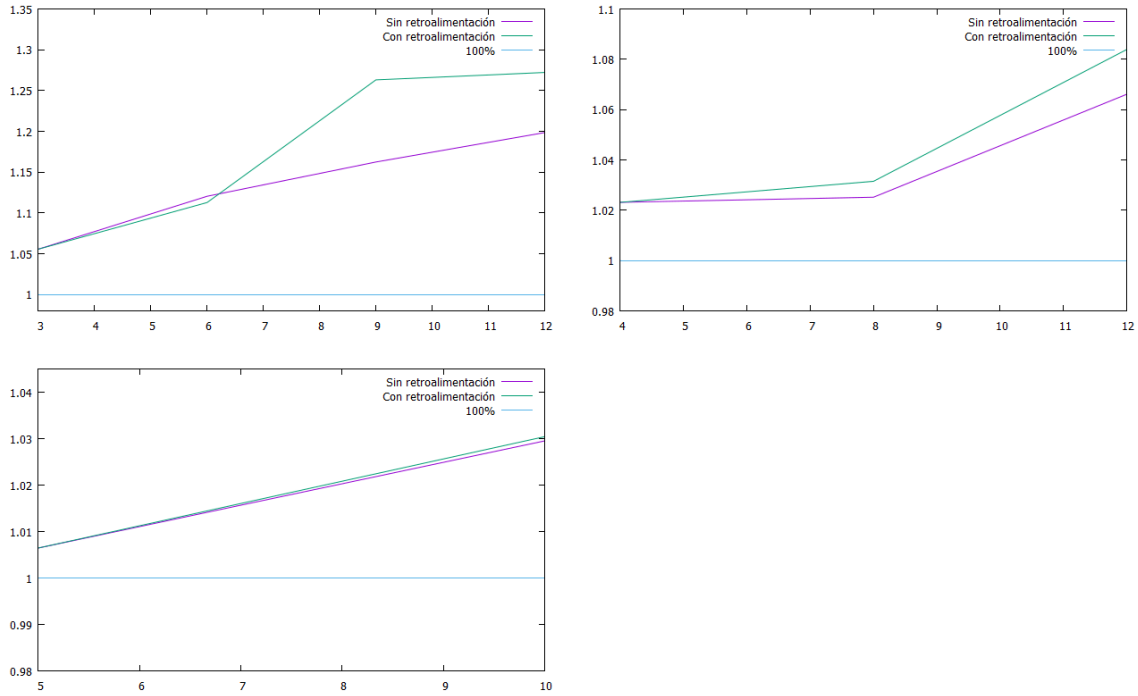
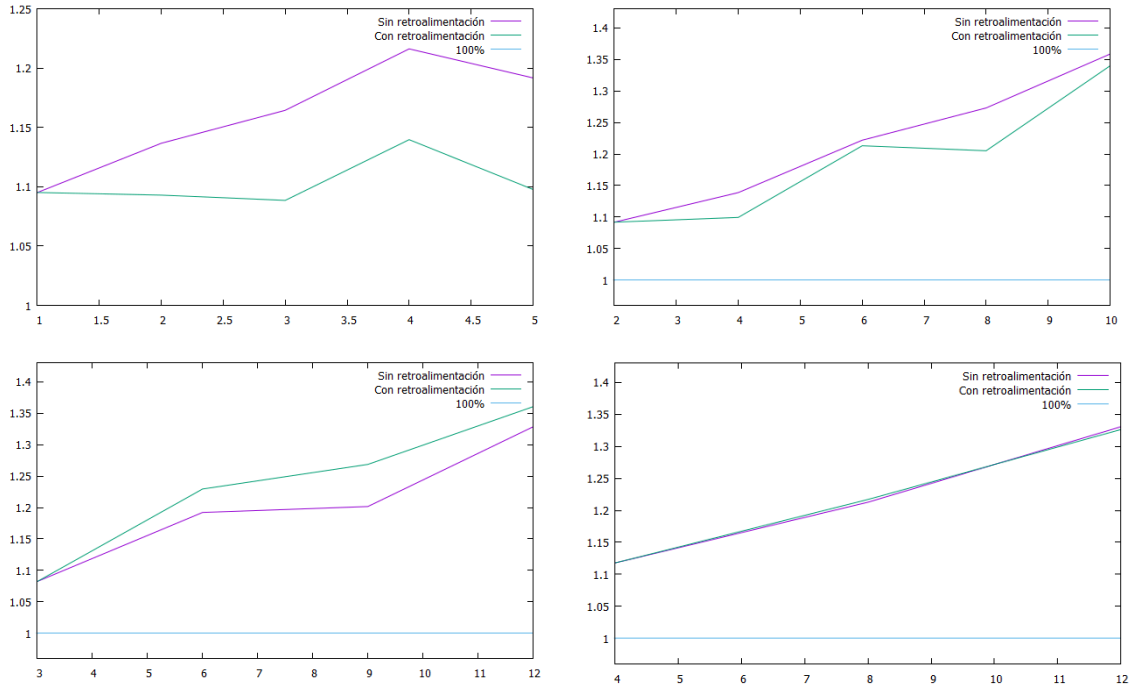


Figura 49: Precisión para 70 años Periodo pronóstico [1-5] años

En los pronósticos para 80 años, donde se puede comprobar donde los valores conseguidos no son bueno, como ocurría con el caso de cohorte, para esta misma edad.



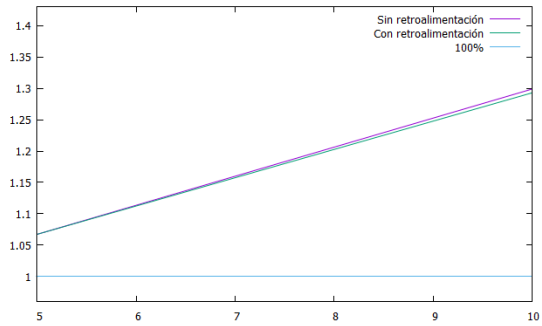


Figura 50: Precisión para 80 años Periodo pronóstico [1-5] años

Por último se muestran los datos pertenecientes a la edad de 90 años, donde se confirma que el error es mucho menor ahora que en el caso con cohorte.

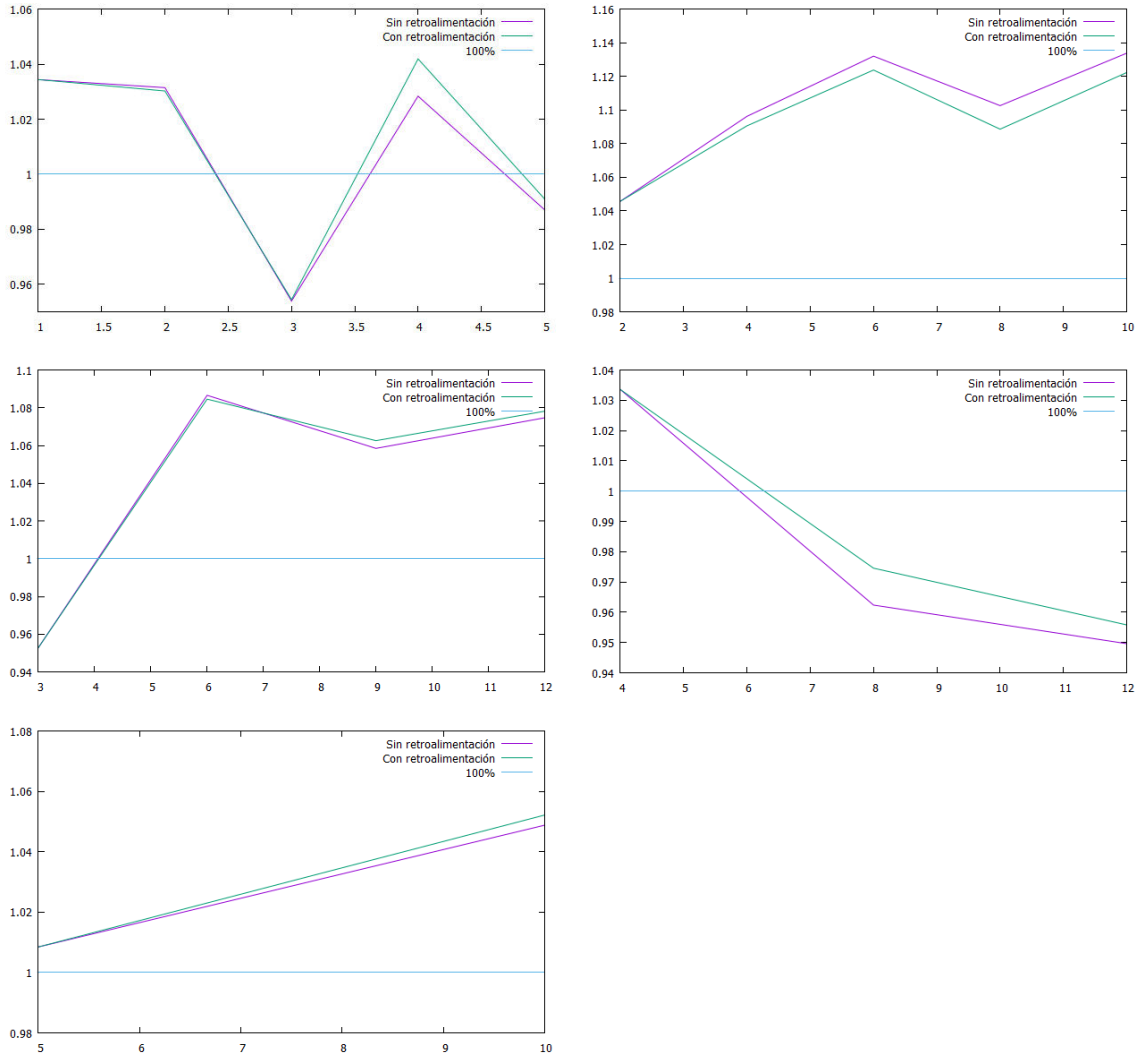


Figura 51: Precisión para 90 años Periodo pronóstico [1-5] años

Si nuevamente como en el caso de cohorte, se realizan los experimentos con un tramo más pequeño de entrenamiento, y en vez de utilizar patrones que van de 1960 al 2000, se utilizan patrones que van de 1985 al 2000, los resultados son los que se pueden observar en la gráfica 52, que realmente no son mejores con un periodo inferior como ocurrió con el pronóstico de tipo cohorte.

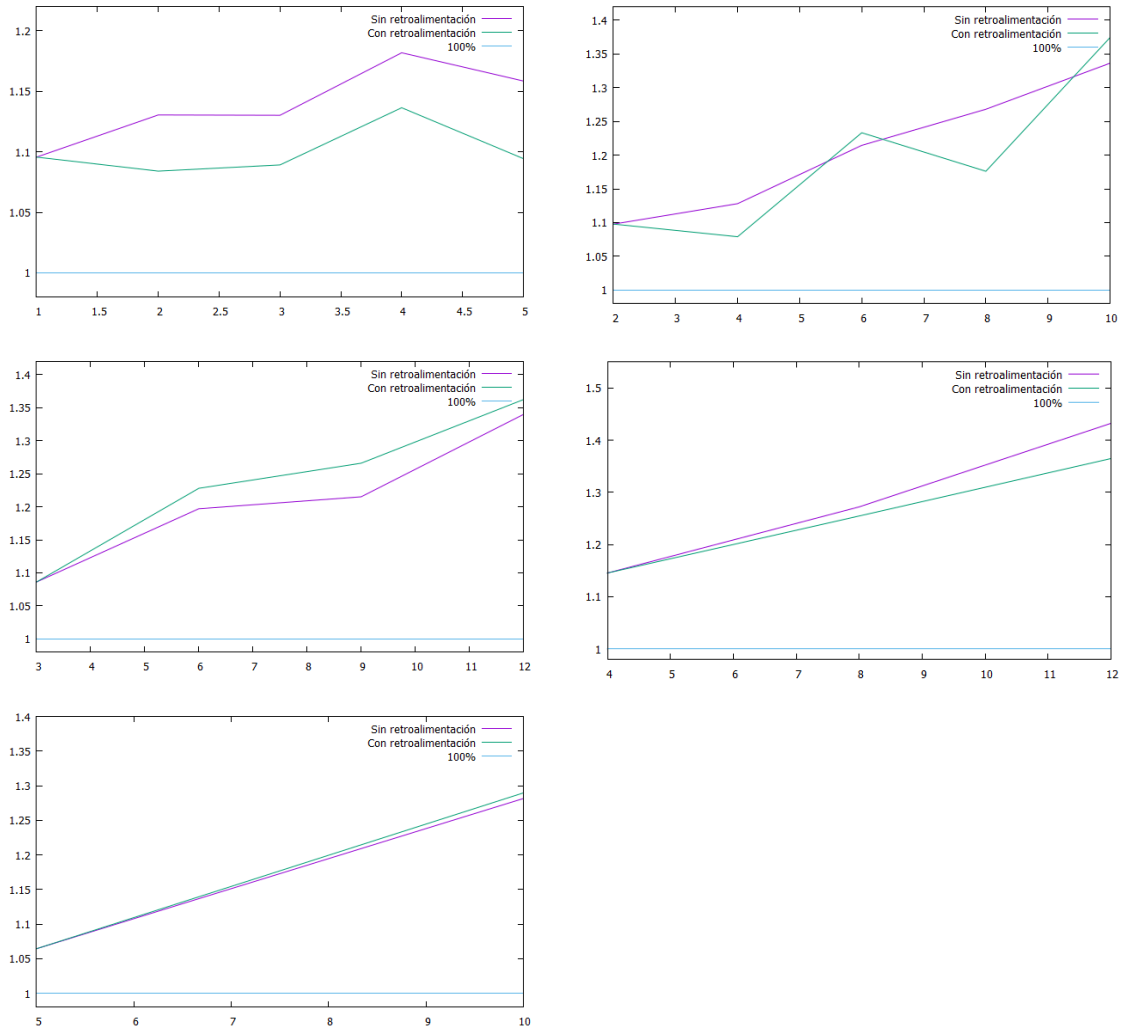


Figura 52: Precisión para 80 años Periodo pronóstico [1-5] años

Si se comprueba la gráfica de los datos que sobre mortalidad (figura 53) se tienen se ve que tiene muchos altibajos, con lo que el problema podría ser que se valide con sólo cuatro elementos, siendo estos elementos los cuatro últimos. En la gráfica 53 se ven los datos y el pronóstico realizado por la red, se puede comprobar que seguirían la hipotética línea marcada por los 4 puntos de validación.

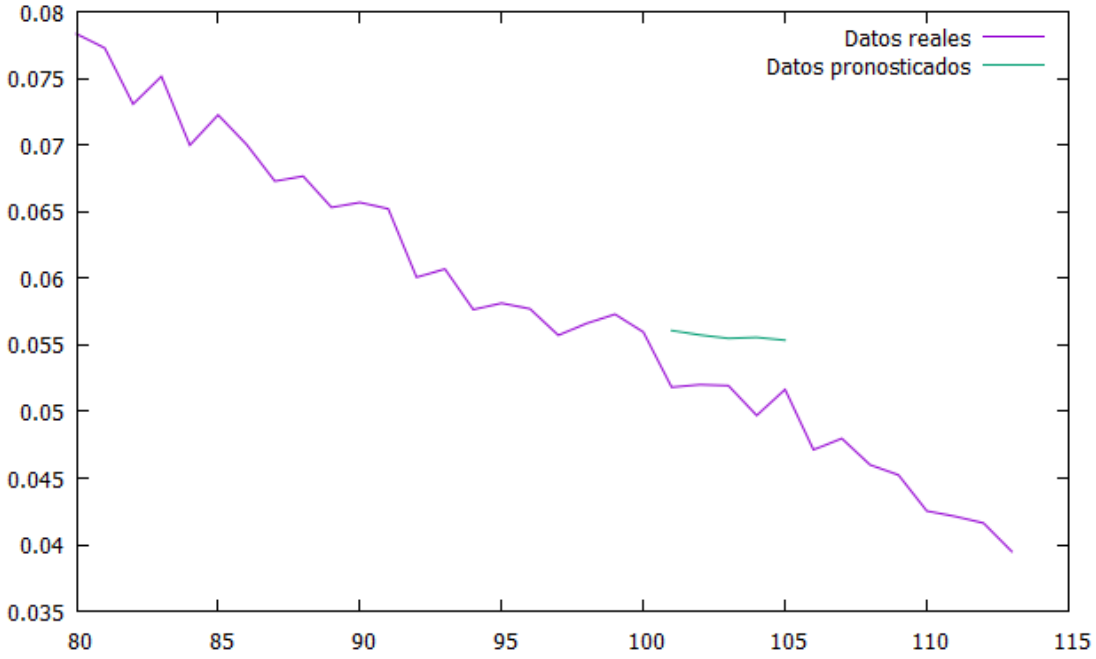


Figura 53: Datos reales frente a pronóstico

6.3 – Conclusiones

A la vista de los resultados, la utilización de las redes de neuronas para la predicción de las tablas de mortalidad, tienen una importante componente de prueba y error ya que la influencia que pueden tener los parámetros en la consecución de una u otra red es muy importante.

Las redes de neuronas no son algo determinista, sino que dependiendo de la inicialización de centroides, desviaciones, y pesos, se dará lugar a unos resultados u otros. Cuando se lanzan varios experimentos, y en cada uno de los experimentos se entrena una serie de redes, de las cuales se escoge la mejor, y las mejores redes de cada experimento devuelven valores muy parecidos entre sí, y además estos valores o pronósticos son muy parecidos a los reales al realizar el testing, más confiados podremos estar de que en efecto esas redes predicen convenientemente los valores futuros.

El problema se presenta cuando cada vez que se lanza un experimento, los pronósticos de la red difieren mucho con los valores de cada uno de los experimentos lanzados. En este caso es necesario comprobar la gráfica de los datos reales, y ver qué es lo que puede estar sucediendo, si se está trabajando con demasiados valores de entrenamiento, o demasiado pocos, y lo mismo con la parte de validación.

Otro problema puede provenir, de que el número de neuronas de la capa oculta sea desproporcionado al número de patrones de entrenamiento. En el caso de este trabajo el número de patrones no era grande, por lo que el número de neuronas no debe ser muy elevado. En general suele ser menor que el número de patrones de entrenamiento, porque en otro caso se puede producir un sobreaprendizaje de estos patrones, y una falta de generalización que afecta a la validación y por tanto al desempeño de la red.

Lo anterior viene referido principalmente al caso de aprendizaje híbrido, ya que al usar el algoritmo K-Medias, cada centroide se posiciona en las coordenadas de cada patrón de entrada, de forma que introducir más neuronas en la capa oculta que patrones de

entrada existen, introduce en general una fuente de ruido, que genera una mayor imprecisión en los pronósticos.

En el caso del aprendizaje totalmente supervisado, que es el usado en la experimentación, las neuronas de las capa oculta se reposicionan con cada patrón de entrenamiento, por lo cual se pueden introducir más neuronas que patrones de entrada existen, de hecho, es posible que en experimentos realizados con rangos de 30 a 50 neuronas en la capa oculta, se pueda mejorar sensiblemente los resultados por ejemplo para la edad de 50 años, edad cuyos pronósticos eran en general algo peores que por ejemplo para 70 años, ya que los rangos óptimos de la parametrización puede variar dependiendo de la edad escogida para pronosticar.

Se ha podido comprobar como para ciertos años de pronósticos para cierta edad, la precisión era muy alta, mientras que para otros era bastante baja, incluso el caso por ejemplo, de que para 60 años el pronóstico sea mejor de 4 en 4 años que de 3 en 3 años.

En general los valores obtenidos para las edades de 60, 70 y 80 años han sido bastante buenos, teniendo en cuenta que los valores en un primer momento para los 80 años (hablando de cohorte) fueron bastante mediocres, se vio como al realizar modificaciones en la parametrización, las redes así generadas devolvían unos buenos valores.

Para el caso del pronóstico de tipo periodo a los 80 años, aunque el tramo de años escogido para entrenar la red influyó positivamente, no fue tan bueno ni mucho menos como en el caso de la red para el tipo cohorte.

Parece que para las edades altas, 80 en adelante, hay mayores problemas para conseguir entrenar una red que de buenos pronósticos. Quizá se debida a la variabilidad de los datos de la tabla. A mayor edad el número de expuestos decrece, por lo que la mortalidad producida incide de forma más aguda en la probabilidad de muerte para esa edad, esto contribuye a una mayor fluctuación de los valores, y por tanto que el pronóstico para estas edades sea más difícil.

No hay que olvidar que la dificultad para obtener pronósticos fiables, en el caso de edades avanzadas, es un hecho, y que por esta razón se han desarrollado modelos específicos como el de *Kannistö.*

El uso por tanto, de un modelo basado en una red de neuronas de base radial, como en el de este trabajo, quizá no sea lo más indicado para ciertos tramos de edad, pero sí que es posible utilizarlas para otros tramos en los que este tipo de red es capaz de pronosticar muy bien. Ciertamente, en aquellos tramos donde el pronóstico sea más pobre, sería más conveniente usar los valores reales de la tabla de mortalidad, ya que la diferencia del valor actual de mortalidad con el que tendrá al año siguiente, podría ser menor que el que la propia red de neuronas determina.

En ocasiones también, se ha dado el hecho de que una red pronosticaba mejor a cinco años que a dos o a tres, habiéndose realizado varios experimentos para cada distinto año de pronóstico, con valores devueltos muy parecidos para uno y otro experimento, cuando se traba de pronosticar a los mismos años.

Lo que se quiere decir, es que cuando se realizan varios experimentos, y cada experimento involucra el entrenamiento de gran cantidad de redes de neuronas, y como resultado de cada experimento se devuelve la catalogada como mejor red de neuronas, atendiendo al error de validación, y esta red provee resultados muy parecidos a otros experimentos para la misma edad y año de pronóstico, existe una consistencia, y por tanto mayor confianza en esos valores devueltos por la red.

Sí a lo anterior le sumamos que existe una consistencia a tener más error para pronóstico a menos años que a más años, se deberían estudiar esos patrones con los que se entrenó la red, para llegar a alguna conclusión de por qué puede estar pasado esto, y modificar los parámetros de entrenamiento y validación de la red.

En general es mejor recoger el primer valor pronosticado por cada tipo de red, es decir, se puede tener una red que pronostica año a año, de tal forma que se podría querer pronosticar con esta red el quinto año, pero en general es mejor que esa red pronostique el primer año, y el segundo que lo pronostique una red entrenada para realizar

pronósticos de dos en dos años, y así para cada uno de los años que se desee pronosticar.

En general como he comentado esto es así, no obstante es cierto que hay casos donde esto no ha ocurrido, de nuevo esto depende de los datos con los que se entrenó, dado que fluctuaciones en estos datos incidirán en los datos que la red devolverá.

6.4 – Trabajos futuros

En este trabajo se ha implementado una red de neuronas de base radial, se podrían estudiar las formas en las que los parámetros de entrenamiento inciden en los resultados, y en cómo sería la mejor forma de escoger estos patrones de entrenamiento, para de esta forma mejorar los pronósticos.

Estudiar la razón de por qué con el método híbrido de entrenamiento, se obtienen peores resultados que con el método totalmente supervisado, quizá la disposición de los centroides sea determinante, o bien las desviaciones sean las que inciden de forma más contundente a los malos resultados. En el método híbrido, se podría utilizar otro algoritmo para la determinación de los centroides, en vez de usar K-Medias.

Se podrían idear nuevas arquitecturas para este tipo de red de neuronas, por ejemplo una arquitectura retroalimentada en el entrenamiento de la red, y ver como esto afecta a la generalización de la red de neuronas y a los pronósticos por ella devueltos.

Se podría implementar una red de tipo NAR o NARX (Nonlinear autoregressive with exogenous inputs) y como varía la capacidad de predicción respecto a la aquí implementada.

También se podría comparar esta red con la de tipo perceptron multicapa, y ver cuál de las dos es más conveniente para este problema en particular. O bien la diferencia en la predicción entre una red de neuronas y los modelos más comúnmente usados en series temporales.

Otra posibilidad es la de comprar, los modelos más exitosos en este ámbito como *Lee Carter, Renshaw-Habermann* o *P-Splines* por citar algunos, con las redes de neuronas entrenadas. Quizá en algunos tramos de edad podrían existir sorpresas.

Otra opción sería idear ciertos refinamientos a la hora de entrenar la red de neuronas, y ver cómo inciden a la hora de su funcionamiento. Por ejemplo se podría una vez entrenada la red con los parámetros en cuestión, realizar un entrenamiento con una razón de aprendizaje muy baja que ajustara mejor los pronósticos, y ver si esto mejora la precisión de forma consistente en la mayoría de los casos probados.

Otra posible experimentación es estudiar los patrones de entrada, y realizar una selección de acuerdo a cierta hipótesis de ciertos patrones, y a la hora de entrenar la red de neuronas, y ver si produce una mejora en precisión.

La parametrización de la red es algo fundamental, y aunque se han escogido unos rangos en los parámetros, para todas las edades con las que se ha experimentado, se podría considerar un distinto rango para una determinada edad. Por ejemplo en el número de neuronas de la capa oculta, ya que los intervalos escogidos lo han sido de acuerdo a lo devuelto por el algoritmo genético para una edad determinada, y estos valores se han utilizado para todas las edades, así pues, mietras que para 60 años el rango ideal estaba entre 10 a 20 neuronas en la capa oculta, quizá fuera mejor ampliar el rango para la edad de 50 años, que como se ha visto se obtenían peores pronósticos.

Al margen del pronóstico de los valores de una tabla de mortalidad, pero relacionado en cierta forma con lo que se persigue de una, un posible trabajo futuro relacionado con el poder clasificador de las redes de neuronas, es por ejemplo entrenar una red que a través del cuestionario sea capaz de determinar, qué personas de la cartera se suponen que van a vivir más que el resto.

De esta forma se podría deducir cuanta repercusión en la mortalidad, respecto de la cartera que se posee, tendrán esas personas, y por tanto si el riesgo de longevidad podría ser elevado.

BIBLIOGRAFÍA

BIBLIOGRAFÍA

[Rodríguez-Pardo, 2014] José Miguel Rodríguez-Pardo del Castillo, Irene Albarrán Lozano, Fernando Ariza Rodríguez,Víctor Manuel Cóbreces Juárez, María Luz Durbán Reguera. *El riesgo de longevidad y su aplicación práctica a Solvencia II Modelos actuariales avanzados para su gestión*, ed Fundación Mapfre.

[Nieto y Vegas, 93] Nieto, U. y Vegas, J. (1993). *Matemática Actuarial. Mapfre, Madrid.*

[Villalón, 1994] Villalón, J. G. (1994). *Manual de Matemáticas Financiero-Actuariales. Fernández Ciudad, S L, Madrid.*

[HMDB] http://www.mortality.org

[Werbos, 74] P.J. Werbos.
Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences.
PhD thesis, Harvard University, Cambridge, MA 02142, August 1974.

[Haykin, 94] Haykin, Simon.
Neural Networks. A Comprehensive Foundation.
Prentice Hall, 1994.

[Kröse & Smagt, 93] Ben J.A. Kröse y P. Patrik van der Smagt.
An introduction to neural networks. 1993.

[Barto, Sutton & Anderson, 83] G. Barto, R. S. Sutton, and C. W. Anderson.
Neuronlike adaptive elements that can solve dificult learning control problems.
IEEE Trans on systems, man and cybernetics, SMC-13:834--846, 1983

[Kohonen, 01] Teuvo Kohonen.
Self-ORganizing Map.
Springer Series in Information Sciences, Vol. 30, Springer, Berlin, Heidelberg, New York, 2001.

[Freeman & Skapura, 93] James A. Freeman / David M. Skapura.
Redes Neuronales. Algoritmos, aplicaciones y técnicas de programación.
Addison-Wesley/Diaz de Santos 1993.

[MCCULLOCH & PITTS , 1943] WARREN S. MCCULLOCH AND WALTER PITTS University of Illinois, College of Medicine, Department of Psychiatry at the Illinois Neuropsychiatric Institute, University of Chicago, Chicago, U.S.A.

[Blum, 92] Adam Blum.
Neural Networks in C++, an objet-oriented framework for building connectionist systems. Wiley, 1992.

[Hecht, 89] Hecht, Nielsen, R. *Theory of the Backpropagation Neural Network.*
Washington, D. C., "Int. Conf. on Neural Networks", Vol.1, 593-605 (1989).

[Hecht, 88] Hecht, Nielsen, R.
Neurocomputing: Picking the Human Brain.
IEEE Spectrum 25(3), March 1988, pp. 36-41

[K.A. De Jong ,1975]. *An analysis of the behaviour of a class of genetic adaptive systems. Tesis doctoral, University of Michigan.*

[HESSER, J., y MÄNNER, R. 1991] *Toward an optimal mutation probability for genetic algorithms*. En:
Proceedings of the first international conference on parallel problem solving from nature (PPSN I),pp. 23-32.

[Quinlan, 91] Quinlan.
Connectionism and Psychology.
Harvester Wheateaf. N.Y. 1991.

[Rumelhart & McClelland, 86] Rumelhart, D.E., McClelland, J.L. & Group, PR.
Parallel Distributed Processing. Explorations in the Microstructure of Cognition.
Cambridge, MA: MIT Press, 1986.

[Widrow, 59] Widrow, B.
Adaptative sampled-data systems, a statistical theory of adaptation.
IRE WESCON Convention Record, part 4, 1959. New York: Institure of Radio Engineers.

[Hebb, 49] Hebb, D.O.
Organization of behavior.
New York: Science Editions, 1949.

[Rosenblatt, 58] Rosenblatt, F.
The perceptron: a probablistic model for information storage and organization in the brain.
J. Andersen & E. Rosenfeld, eds, `Neurocomputing: foundations of research', Bradford books, MIT Press, 1958, Cambridge, Mass., chapter 8.

[Widrow, 62] Widrow, B.
Generalization and information storage in networks of <<Adaline>> In Self-Organizing Systems.
Eds. M.C. Yovits, G.T. Jacobi et G.D. Goldstein, Washington, 435-461, 1962.

[Minsky & Papert, 69] Minsky, M.L. & Papert.
Perceptrons.
Cambridge, MA: MIT Press, 1969.

[Hoplfield, 82] Hoplfield, J.L.
Neural networks and physical systems with emergent collective computational abilities.
Proceedings of the National Academy of Science USA, 1982, 2554-2558.

[Grossberg, 87] Grossberg, S.
Competitive learning: from interactive activation to adaptative resonance.
Cognitive Science, 11, 1987, 23-63.

[Narendra & Thathchar, 74] Narendra, K. S. and Thathchar.
Learning Automata - A Survey.
IEEE Trans. on SMC, Vol. 14, 1974, pp. 323-334.

[Chirungrueng & Séquin, 95] Chirungrueng, C. and Séquin C.
Optimal Adaptive K-Means Algorithm with Dynamic Adjustment of Learning Rate.
IEEE Transactions On Neuronal Networks, Vol6, NO. 1, January 1995.

[Gersho, 79] Gersho, A.
Asymptotically optimal block quantization.
IEEE Trans. inform. Theory, vol. IT-25, no. 4, pp.373-380, 1979.

[Valls, Galván & Molina, 00] Galván, I.M., Valls, J.M. y Molina, J.M.
Sistema Multiagente para el diseño de Redes de Neuronas de Base Radial óptimas.
Revista Iberoamericana de Inteligencia Artificial, Nº 10 (2000), pp 18-25.

[Platt, 91] Platt, J.
A Resource-Allocating Networks for Function Implementation.
Neural Computation 3, 213-225, 1991.

[Moody & Darken, 89] Moody J.E. and Darken C.J.
Fast Learning in Networks of Locally-Tuned Processing Units.
Neural Computation 1, 281-294, 1989.

[Poggio & Girosi, 90] Poggio T. and Girosi F.
Networks for approximation and learning.
Proceedings of the IEEE, 78, 1481-1497, 1990.