



núcleo de ingeniería biomédica
facultades de medicina e ingeniería



universidad de la república

Redes neuronales artificiales

Andres Nacelle,

Monografía vinculada con la conferencia del Prof. Dr. Eduardo Mizraji sobre: "Las redes neuronales: de la biología a los algoritmos de clasificación" de la fecha 9 de junio 2009

¹ Núcleo de Ingeniería Biomédica - Facultades de Medicina e Ingeniería - Universidad de la República, Uruguay

Email: Andres Nacelle - andresnacelle[en]hotmail[.]com;

Resumen

Fundamento - En este trabajo se estudian las bases que permiten comprender la teoría detrás de las redes neuronales artificiales desde su motivación biológica hasta su implementación. Inicialmente se analiza el cerebro y la neurona como inspiración y modelo para luego explicar los componentes de una red neuronal artificial y en función de estos clasificarlas. Por último se analizan diferentes opciones de implementación y se contrastan los modelos analógicos y digitales.

Metodología - La elaboración de este documento implicó el estudio y contraste de material de diversas fuentes y con varios años de diferencia, siendo más difícil obtener material del comienzo de los 70's y mediados de los 90's, debido al estancamiento de los avances en los modelos vigentes en cada época. Sin duda para esta tarea el acceso a Internet como herramienta fue imprescindible.

Resultados - Se establece al perceptrón como un modelo simple y claro pero con la gran limitante que solo permite trabajar problemas linealmente separables, lo que significó una época gris en la investigación. Esta situación se revierte con el advenimiento de las redes multicapa que implementaban nuevos algoritmos de aprendizaje y la posibilidad del modelado no lineal, los cuales a pesar de requerir de mucho tiempo de entrenamiento y grandes cantidades de datos etiquetados han generado excelentes herramientas que se emplean en todos los campos.

1. Introducción

El campo de las redes neuronales artificiales fue inspirado por los estudios en redes neuronales reales que conforman el cerebro, en un intento por resolver problemas mediante el empleo del conocimiento obtenido de casos similares resueltos en el pasado. No por eso su conducta es igual a la del cerebro, pero si se busca emular ciertas cualidades destacables de este, las cuales no se alcanzan con computadoras estándar.

2. El modelo Biológico

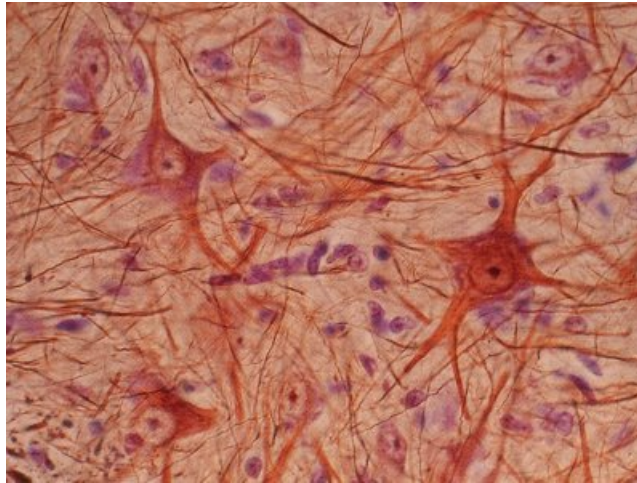


Figura 1: Redes neuronales, en el cerebro estas son altamente complejas.

El cerebro como sistema de procesamiento de información posee en el orden del billón de células nerviosas, donde cada neurona esta conectada a otras neuronas a través de unas 10000 sinapsis. Algunas de sus propiedades de interés son:

1. La capacidad, que poseen partes sanas, de aprender a encargarse de tareas anteriormente realizadas por partes ahora dañadas.
2. Su capacidad de aprendel de la experiencia (lo cual es una propiedad clave).
3. Realiza de forma muy eficiente cálculos masivamente paralelos, como la percepción visual que ocurre en menos de 100 milisegundos.
4. Es responsable de nuestra memoria, intelecto y conciencia de existencia .

El elemento constructivo y unidad básica procesadora de esta red neuronal es la neurona. Su estructura, como se muestra en la Figura 2, es básicamente un cuerpo celular, o soma, que alberga el núcleo y del cual se extienden las dendritas (terminales receptoras) y el axón (elemento transmisor). La unión entre el axón y la dendrita de dos neuronas se llama sinapsis. La comunicación entre neuronas se basa en la activación de terminales bio-química en el extremo del axón, que liberan neurotransmisores que son captados por la dendrita. La experiencia en cierta tarea o estímulo produce que las sinapsis correspondientes se refuercen de forma exitatoria o inhibitoria.

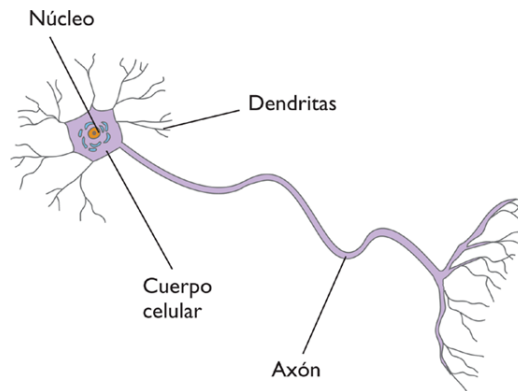


Figura 2: Esquema elemental de una neurona.

3. Comparacion entre cerebro y computadora

Cuadro 1: cerebro vs computador.

	Unidades procesadoras	Tamaño de las unidades	Consumo	Frecuencia de trabajo *	Aprendizaje
Cerebro	10^{14} sinapsis	10^{-6} m	30 W	Cientos de Hz	Si
Computador	10^8 transistores	10^{-6} m	30 W	Unidades de GHz	Poco

* Los problemas de índole aritméticos no hay duda que son mejor manejados por los computadores, sin embargo hay tareas como el reconocimiento de patrones y comprensión del lenguaje en las que el cerebro es mejor en términos de velocidad de procesamiento.

4. Modelo de redes neuronales artificiales

Las redes neuronales artificiales son motivadas por ciertas cualidades de su modelo real, por lo cual el desafío es producir un modelo que tenga:

1. Una estructura de procesamiento distribuida y paralela (opuestamente al CPU de una computadora).
2. Alto grado de conexión entre las unidades básicas.
3. Conexiones modificables en función de la experiencia.
4. Un proceso de aprendizaje constante y de ser posible uno no supervisado
5. Aprendizaje basado en información local.
6. Robustez en la performance si algunas unidades son removidas.

5. Elementos de una red neuronal artificial

El elemento fundamental de una red neuronal es la "neurona", también denominada "nodo", cuyo modelo puede verse en la Figura 3. En ella, la suma de las n entradas x_j de la neurona i , ponderadas con los pesos sinápticos w_{ij} , genera la entrada ponderada total o "potencial postsináptico" de la neurona i . Los pesos sinápticos w_{ij} miden la intensidad de la interacción entre las dos neuronas que están conectadas por el enlace. Posteriormente, se aplica una función de activación o transferencia (f) a la diferencia entre el "potencial postsináptico" y el umbral θ_i , obteniéndose la salida de la neurona (y_i).

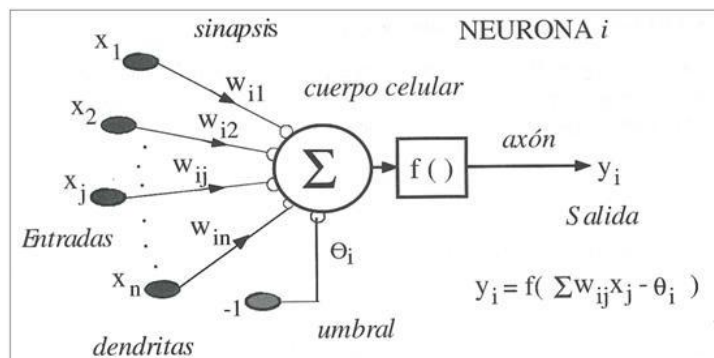


Figura 3: Modelo para neurona artificial o Perceptrón simple

Donde $y_i(t) = f(\sum_{j=1}^n w_{ij}x_j - \theta_i)$ es el modelo matemático mas elemental. La función de activación $f()$ se suele considerar determinista, y en la mayor parte de los modelos es monótona creciente y continua. La forma $y_i(t) = f(x)$ de las funciones de activación más empleadas en las redes neuronales artificiales se muestra en la siguiente tabla, donde x representa el potencial postsináptico e y el estado de activación. Este modelo de neurona, también llamado Perceptrón, fue el primer modelo de red neuronal (1957) que le abriría las puertas a las primeras investigaciones. En 1969 se demuestra matemáticamente que solo puede implementar funciones linealmente separables, lo cual trae consigo siete años de estancamiento.

	Función	Rango	Gráfica		Función	Rango	Gráfica
Identidad	$y = x$	$[-\infty, +\infty]$		Sigmoidea	$y = \frac{1}{1 + e^{-x}}$ $y = \tanh(x)$	$[0, +1]$ $[-1, +1]$	
Escalón	$y = \text{sign}(x)$ $y = H(x)$	$\{-1, +1\}$ $\{0, +1\}$		Gaussiana	$y = Ae^{-Bx^2}$	$[0, +1]$	
Lineal a tramos	$y = \begin{cases} -1, & \text{si } x < -l \\ x, & \text{si } -l \leq x \leq +l \\ +1, & \text{si } x > +l \end{cases}$	$[-1, +1]$		Sinusoidal	$y = A \sin(\omega x + \varphi)$	$[-1, +1]$	

Figura 4: Algunas de las funciones de activación mas empleadas.

Un sistema eléctrico discreto que implementa en hardware el modelo matemático descripto es:

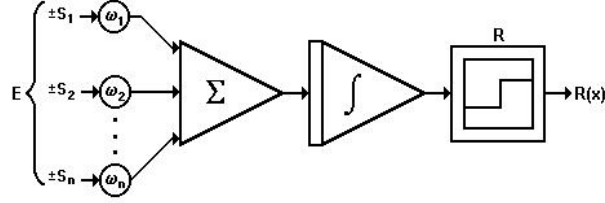


Figura 5: Modelo electrónico con componentes discretos.

Las neuronas artificiales se pueden estructurar en capas interconectando cada neurona con la siguiente con un peso asociado, pudiendo haber o no, interacción entre neuronas de una misma capa. De este modo las capas se pueden apilar dando lugar a complicadas redes con capa de entrada, capa de salida y capas ocultas o intermedias. Cualquiera de estas redes puede representarse por una matriz de pesos con todos los pesos de la red. Si w_{ji} es positivo indica que la relación entre las neuronas es excitadora, es decir, siempre que la neurona i esté activada, la neurona j recibirá una señal que tenderá a activarla. Si w_{ji} es negativo, la sinapsis será inhibitoria. En este caso si i está activada, enviará una señal que desactivará a j . Finalmente si w_{ji} es 0 se supone que no hay conexión entre ambas. Estas estructuras más complejas que el Perceptrón fueron las que provocaron el resurgimiento en 1986 de una fuerte línea de investigación, pues la concatenación de niveles permitía modelar no linealidades y esto significa poder aplicar dichos modelos a la solución de un mundo nuevo de problemas.

En función de como se realiza la interconexión entre neuronas se pueden tener las diferentes arquitecturas:

1. Red monocapa (figura 6.a.), se corresponde con la red neuronal más sencilla ya que se tiene una capa de neuronas que proyectan las entradas a una capa de neuronas de salida donde se realizan los diferentes cálculos.
2. Red multicapa (figura 6.b.), es una generalización de la anterior, existiendo un conjunto de capas intermedias entre la capa de entrada y la de salida (capas ocultas). Este tipo de red puede estar total o parcialmente conectada.
3. Red recurrente (figura 6.c.), este tipo de red se diferencia de las anteriores en la existencia de lazos de realimentación en la red. Estos lazos pueden ser entre neuronas de diferentes capas, neuronas de la misma capa o, entre una misma neurona. Esta estructura la hace especialmente adecuada para estudiar la dinámica de los sistemas no lineales.

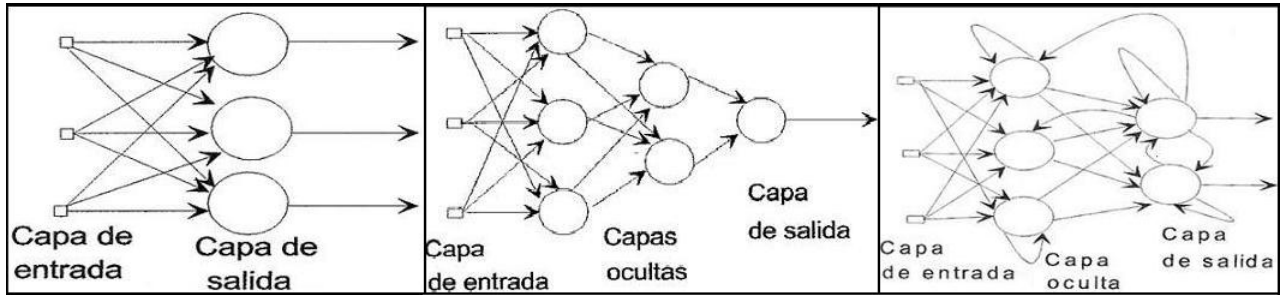


Figura 6: Posibles topologías: izq 6.a Red monocapa, centr 6.b Red multicapa, der 6.c Red recurrente

Un ultimo elemento en las redes neuronales artificiales es la regla de aprendizaje. Este concepto es sustancialmente la modificación de la respuesta producido por la interacción con el medio y de esta forma establecer nuevas respuestas que se adaptan a estímulos externos. En el caso del cerebro estas cualidad se alberga en las sinapsi en el caso de las redes neuronales, el conocimiento esta en los pesos de las conexiones. Por esto, todo mecanismo de aprendizaje demanda modificaciones en dichos pesos, las cuales se pueden hacer a dos niveles:

1. A través del modelado de las sinápsis; que consiste en modificar los pesos sinápticos siguiendo una cierta regla de aprendizaje, construida normalmente a partir de la optimización de una función de error, que mide la eficacia actual de la operación de la red. Si denominamos $w_{ij}(t)$ al peso que conecta la neurona presináptica j con la postsináptica i en la iteración t , el algoritmo de aprendizaje, en función de las señales que en el instante t llegan procedentes del entorno, proporcionará el valor $\Delta w_{ij}(t)$ que da la modificación que se debe incorporar en dicho peso, el cual quedará actualizado de la forma: $\Delta w_{ij}(t+1) = w_{ij}(t) + \Delta w_{ij}(t)$ El proceso de aprendizaje es usualmente iterativo, actualizándose los pesos de la manera anterior, una y otra vez, hasta que la red neuronal alcanza el rendimiento deseado.
2. A través de la creación o destrucción de neuronas; en el cual se lleva a cabo una modificación de la propia arquitectura de la red.

De este modo algunos de los tipos de aprendizaje mas relevantes que se pueden distinguir son:

1. Aprendizaje supervisado. En el aprendizaje supervisado se presenta a la red un conjunto de patrones, junto con la salida deseada u objetivo, e iterativamente ésta ajusta sus pesos hasta que la salida tiende a ser la deseada, utilizando para ello información detallada del error que se comete en cada paso. De este modo, la red es capaz de estimar relaciones entrada/salida sin necesidad de proponer una cierta forma funcional de partida. Es decir, si $E[w]$ es una función que representa el error esperado de la operación de la red, expresado en función de sus pesos sinápticos w , se pretende estimar una cierta función multivariantes $f : R^m \rightarrow R^n$ a partir de muestras (x, y) tomadas aleatoriamente por medio de la minimización iterativa de $E[w]$ mediante aproximación estocástica (las técnicas de aproximación estocástica estiman valores esperados a partir de cantidades aleatorias observadas).
2. Aprendizaje no supervisado o auto organizado. El aprendizaje no supervisado se puede describir genéricamente como la estimación de la función densidad de probabilidad $p(x)$ que describe la distribución de patrones $x \in R^n$ (espacio de entrada). En este tipo de aprendizaje se presentan a la red multitud de patrones sin adjuntar la respuesta que deseamos. La red, por medio de la regla de aprendizaje, estima $p(x)$, a partir de lo cual podemos reconocer regularidades en el conjunto de entradas, extraer rasgos o agrupar patrones según su similitud
3. Aprendizaje híbrido. En este caso existen en la red los dos tipos de aprendizaje básicos, supervisado y autoorganizado, normalmente en distintas capas de neuronas.
4. Aprendizaje reforzado. Se sitúa a medio camino entre al aprendizaje supervisado y el autoorganizado. Como en el primero, se emplea información sobre el error cometido, pero en este caso existe una única señal de error, que representa un índice global del rendimiento de la red (solamente le indicamos lo bien o lo mal que está actuando). Como en el caso del no supervisado, no se suministra explícitamente la salida deseada.

Los mecanismos de aprendizaje y determinación de pesos son actualmente unos de los puntos críticos en la investigación de redes neuronales. Entrenar una red de seis o siete capas es una tarea que lleva mucho tiempo y requiere de muchos datos conocidos y etiquetados. Esto llevo a un nuevo declive en las investigaciones a mediados de 1990 en cuanto a modelos. Sin embargo el avance en las técnicas de microelectrónica permitió avanzar en su implementación para tareas específicas. Actualmente un nuevo modelo de aprendizaje se está ensayando por el equipo de Geoffrey Hinton (principal responsable del algoritmo de backpropagation) que promete superar estos problemas entrenando capas de a una y con una cantidad mínima datos etiquetados.

6. Implementación de redes neuronales artificiales

Las limitaciones de los perceptrones para la implementación de problemas llevo a las redes neuronales artificiales a una de sus épocas mas oscuras, de las cuales salieron con los modelos de multicapas capaces de aprender funciones no lineales de gran complejidad, mediante algoritmos de aprendizaje como el de back-propagation. Otro elemento notable que devolvio interés en el área de las redes neuronales fue el desarrollo de la tecnología de VLSI (very large scale integration), sistema que permitía programar microchips con un alto grado de paralelismo. A partir de ahí y con los avances venideros en el área de la informática las implementaciones de redes neuronales encontraron varios nichos.

1. Simulación software de redes neuronales artificiales. Es la implementación de una red neuronal en un ordenador de propósito general en forma de aplicación. Claro esta que por una cuestión de costos y accesibilidad es un modo de implementación natural, pero es necesario ser conciente de las serias limitaciones de este método, pues se busca reproducir el alto paralelismo de las redes neuronales en un procesador secuencial. Por ejemplo, un problema que precise una red de 100 neuronas (sistema relativamente pequeño), incluiría del orden de 10000 conexiones o pesos sinápticos. Si suponemos que un microprocesador realiza alrededor de 10 millones de MIPS, podría ejecutar aproximadamente 1000 fases de recuerdo por segundo. Si el tiempo de respuesta necesario es del orden de milisegundo, la simulación en un ordenador convencional resultará suficiente. Sin embargo, si la aplicación requiere 300 neuronas, la red contendrá 90000 conexiones, con lo que sólo podrá realizar 100 fases de recuerdo por segundo. Así para un problema cuya resolución precise miles o millones de neuronas, como puede ser una tarea de visión o un tratamiento de imágenes (256x256 pixels, cada pixel representado por una neurona), la simulación sobre un ordenador convencional puede resultar inoperante y se hace necesaria una solución distinta, en la que el paralelismo de cálculo sea manifiesto.
2. Emulación de hardware de redes neuronales artificiales. Con esto se hace referencia al uso de elementos de procesamiento con cierto grado de calculo en paralelo, como DSP's, diseñados para el trabajo con redes neuronales colocados en paralelo. Es un punto intermedio entre el software y la implementación en hardware. Se suelen emplear periféricos coprocesadores que mediante un software adecuado, permita aprovechar las capacidades de calculo en paralelo y así acelerar las simulaciones de las redes. Estos neurocomputadores y neurochips son una solución un tanto costosa para el bajo aprovechamiento de las redes neuronales que se obtiene .
3. Implementación en hardware. Es la construcción física de un sistema cuya arquitectura implmente lo mas fiel posible la estructura de las redes neuronales. Es por esto que el sistema deberá soportar un gran número de conexiones, y un flujo elevado de comunicaciones de datos. Estas circunstancias limitan considerablemente el tamaño de la red neuronal que puede ser implementada. Este tipo de realización está especialmente indicado para aplicaciones especiales donde se precisa un alto rendimiento, ya que el diseño se realiza para un cierto modelo de red neuronal y para una aplicación concreta. La realización electrónica se lleva a cabo mediante tecnologías microelectrónicas, en forma de ASIC(Application Specific Integrated Circuit), o mediante FPGA (Field Programmable Gate Arrays). Aunque el diseño de ASIC es lo más habitual, en la actualidad la realización mediante FPGA presenta la ventaja de un periodo de desarrollo más corto. Aunque su densidad de puertas es menor, son fácil mente programables por el usuario mediante herramientas de CAD que permiten captura esquemática, síntesis y simulación a partir de una descripción en un HDL (Hardware Description Language). Las FPGA resultan interesantes para el desarrollo rápido y económico de prototipos o series pequeñas, aunque no permiten diseños tan complejos como los ASIC .

Las especificaciones de algunos neurocomputadores señalan tener conjuntos de procesadores conectados entre si de forma concurrente, como el MARK IV formado por un conjunto de procesadores motorola 68020 asociados cada uno a un coprocesador matemático. Es capaz de simular 25600 neuronas y 5500000 sinapsis, con una velocidad de 5000000 de interconexiones por segundo. Mientras que un neurochip (también dispositivos que emulan redes) como el N64 de Intel posee 64 neuronas, 10000 sinapsis y una vel de procesamiento de

2500 conexiones por segundo (hay que tener en cuenta que las interconexiones en este caso son en el silicio del chip).

7. Implementación analógica frente a digital

Como fue visto anteriormente la implementación de una neurona mediante dispositivos analógicos es simple, pues su funcionamiento se basa en las leyes físicas que rigen la electrónica y esto mismo introduce no lineales que permiten la realización de funciones de activación del tipo sigmoideal. Además su velocidad de procesamiento y nivel de conectividad es mayor que el de los sistemas digitales. Pero por otro lado, este circuito posee grandes limitaciones en cuanto a adaptabilidad pues solo puede implementar una única función de activación y la modificación de sus pesos es complicada pues es necesario remover las resistencias. Sin contar el gran consumo de este sistema, gran superficie ocupada y mayor susceptibilidad frente al ruido y variaciones en la tensión de alimentación que los circuitos digitales. Restan dos problemas mas que son la poca precisión (en general no mas de 8 bits) que imposibilita la aplicación de ciertos algoritmos (como el BackPropagation) y la no escalabilidad de la tecnología.

Los sistemas digitales se basan en el almacenamiento de los pesos de las sinapsis en registros y su procesamiento mediante estructuras de calculo como las ALU's y multiplicadores. La integración digital en un neurochip otorga flexibilidad en el diseño, la posibilidad de integrar el modulo de aprendizaje y la realización de diseños de gran precisión. La tecnología digital ofrece la posibilidad de alcanzar los rangos dinámico manejados por algoritmos complicados y de muchas etapas. Otra de las grandes ventajas del diseño de circuitos digitales VLSI es el acceso a herramientas de diseño que permiten un manejo modular del proyecto y una síntesis automática de la descripción hardware del circuito a construir otorgando así gran escalabilidad y adaptabilidad frente a nuevas tecnologías disponibles. Esto también trae consigo la reducción de tiempo de desarrollo y con el los costos. Claro esta que la desventajas de la implementación digital radica en el alto consumo de área de silicio y como ya se dijo la velocidad de ejecución inferior a la analógica.

8. Conclusiones

El campo de las redes neuronales artificiales es un área joven del conocimiento científico y a pesar de ello apenas pude alcanzar una aproximación general de lo que este tipo de trabajo abarca y las posibilidades que ofrece. Sin duda es una herramienta sumamente potente que aun no se ha encontrado la forma mas eficiente de explotar pues por ejemplo incluso con las grandes computadoras de hoy en día lleva bastante tiempo entrenar una red compleja capaz digamos de reconocer letras, incluso empleando los algoritmos mas avanzados de backpropagation. Pero en ese recorrido se han producido grandes herramientas en diversas disciplinas desde la biología permitiendo aprender mas acerca del cerebro a través de estudiar el comportamiento de redes cuya topología es conocida, o la medicina con analizadores de habla para ayuda de audición de sordos profundos, hasta clasificación de señales de radar y optimización del uso de recursos escasos, pasando por las diversas áreas de la industria, meteorología, manufacturación, finanzas, etc.

Referencias

1. <http://www.youtube.com/watch?v=6ixqKw7uK6o&hl=es>
2. <http://video.google.com/videoplay?docid=6294502841192577731>
3. <http://www.youtube.com/watch?v=AyzOUbkUf3M>
4. <http://nptel.iitm.ac.in/video.php?courseId=1080&p=4>
5. <http://www.youtube.com/watch?v=cM1CF7gUm-U>

6. <http://proton.ucting.udg.mx/posgrado/cursos/idc/pdf/idc/60.pdf>
7. <http://ibersid.eu/ojs/index.php/scire/article/viewFile/1036/1018>
8. <http://www.monografias.com/trabajos12/redneuro/redneuro.shtml>
9. <http://alojamientos.us.es/gtocom/pid/pid10/RedesNeuronales.htm>
10. http://www.uam.es/personal_pdi/psicologia/adarraga/studs/Xerxes/Pag5.html
11. <http://personal5.iddeo.es/winrmute/ia/neuronal.htm>
12. <http://electronica.com.mx/neural/informacion/perceptron.html>
13. <http://www.iberchip.org/IX/Articles/PAP-006.pdf>
14. <http://www.gc.ssr.upm.es/inves/neural/ann2/concepts/biotype.htm>
15. <http://www.e-torredebabel.com/Psicologia/Conexionismo/Conexionismo-Indice.htm>
16. Anales de mecánica y electricidad. Noviembre-Diciembre 2002, pag 28