Capítulo 2. Las Redes Neuronales Artificiales

Capitulo 2. Las Redes Neuronales Artificiales

2.1 Definición Redes Neuronales Artificiales

El construir una computadora que sea capaz de aprender, y de entender el significado de las formas en imágenes visuales, o incluso distinguir entre distintas clases de objetos similares son parte de la problemática a la que se enfrentan los que diseñan computadoras, los ingenieros y los programadores [Freeman & Skapura, 1991].

La incapacidad de la generación actual de computadoras para interpretar el mundo en general no indica, sin embargo, que éstas sean completamente inadecuadas. Generalmente los problemas se presentan cuando tratamos de resolver problemas que involucran un procesamiento en paralelo, utilizando una herramienta de tipo secuencial, la computadora. Uno de los problemas que implican un tipo de procesamiento como el mencionado anteriormente, es el reconocimiento visual de imágenes. Para una computadora el reconocer imágenes aún muy diferentes, es una tarea sumamente difícil, lo que en el caso de los humanos es algo relativamente sencillo. Esto se debe a que los sistemas biológicos poseen una arquitectura distinta (paralelismo masivo) a la de una computadora moderna. Por esta razón es que se han tratado de simular algunas características de la fisiología del cerebro humano para elaborar nuevos procesos de elaboramiento [Freeman & Skapura, 1991].

Podemos definir a una Red Neuronal Artificial (RNA en adelante) como modelos matemáticos inspirados en sistemas biológicos, adaptados y simulados en computadoras convencionales [Lara, 1998]. Las RNAs están inspiradas en el sistema biológico natural. Como es conocido, en este sistema la neurona es la unidad de procesamiento, y aunque las RNAs sean mucho menos complejas que una red neuronal biológica, también realizan cálculos complejos para procesar información.

2.1.1 La computación convencional y la biológica

La computación convencional se caracteriza por el desarrollo de una formación matemática del problema, el desarrollo de un algoritmo para implementar una solución, la codificación del mismo para una máquina específica y por último la ejecución de dicho código. Como se ha observado, este tipo de procesamiento es muy exitoso para resolver modelos matemáticos complejos y de simulación, para realizar tareas repetitivas, rápidas y bien definidas. Sin embargo, cuando éste se lleva a otros ámbitos computacionales, se muestra incapaz de resolver eficientemente problemas de reconocimiento de imágenes, de voz, y de entendimiento de lenguaje natural. También resulta ineficiente en problemas de percepción, adaptación y aprendizaje [Lara, 1998].

Por otro lado, la computación biológica (derivada del procesamiento en sistemas biológicos) se caracteriza por ser masivamente paralela, adaptativa, lenta, altamente interconectada y tolerante al ruido en el medio ambiente y en sus componentes.

De acuerdo con la prensa, publicaciones, y conferencias, las redes neuronales (como parte de la computación biológica) han tenido aplicaciones en el área de procesamiento de imágenes y visión computacional [Schalkoff, 1997], específicamente en el análisis de segmentación. Es por ello que en este proyecto uno de los algoritmos a probar para segmentar palabras estará basado en RNAs o la combinación de estas con otra técnica desarrollada .

2.1.2 Definición y descripción de una Red Neuronal Artificial

Una Red Neuronal Artificial es una estructura compuesta de un número de unidades interconectadas (neuronas artificiales). Cada unidad posee una característica de entrada/salida e implementa una computación local o función. La salida de cualquier unidad esta determinada por su característica de entrada/salida, su interconexión con otras unidades, y (posiblemente) de sus entradas externas. Sin embargo es posible un "trabajo a mano", la red desarrolla usualmente una funcionalidad general a través de una o más formas de entrenamiento [Schalkoff, 1997].

El cerebro humano contiene más de 100 billones de elementos de procesos llamados neuronas, que se comunican a través de conexiones llamadas sinapses. Cada neurona esta compuesta por tres partes fundamentales: el cuerpo, dendritas y axón. El cuerpo en su capa externa tiene la capacidad única de generar impulsos nerviosos. Las dendritas que son como las ramas que salen del cuerpo, poseen algunas conexiones sinápticas en donde se reciben señales que generalmente vienen de otros axones. El axón

se encarga de activar o inhibir otras neuronas las cuales a su vez son activadas por cientos o miles de otras neuronas.

El funcionamiento de un neuron artificial esta basado en este diseno. Basicamente consiste en aplicar un conjunto de entradas, cada una representando la salida de otro neuron, o una entrada del medio externo, realizar una suma ponderada con estos valores y "filtrar" este valor con una funcion como se puede observar en la figura 2.1 [Gomez, 1999].

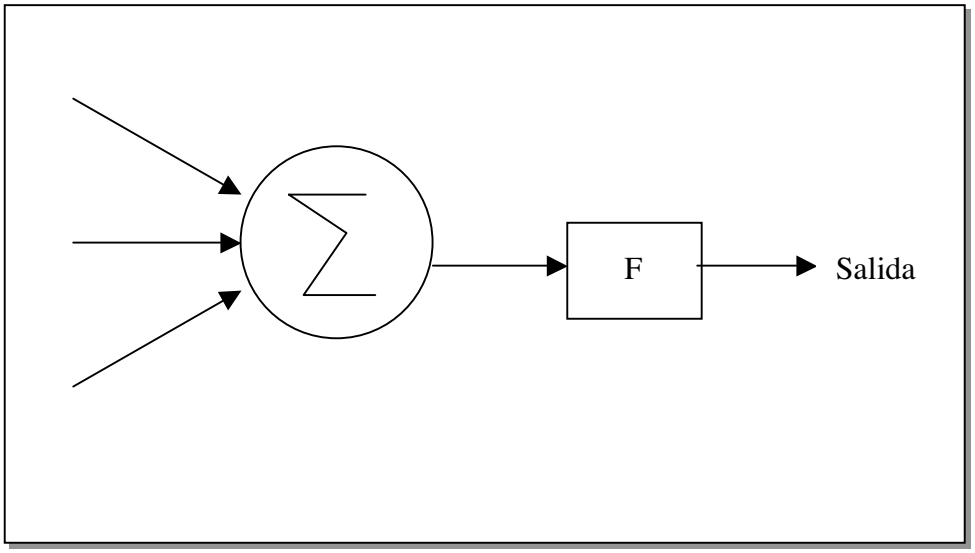

Figura 2.1 Calculo de salida de un neuron artificial [Gomez, 1999]

Cada neuron artificial recibe un vector X de entrada que corresponde a todas aquellas senales que llegan a la sinapsis. Cada una de estas senales se multiplica por un peso que tiene asociado $W_1, W_2, W_3 \ldots W_n$. Al conjunto de pesos se le denomina vector W. Cada peso representa la "intensidad" o fuerza de conexion de una sinapsis en un neuron

biológico. Los resultados de éstas multiplicaciones se suman. Esta sumatoria simula vagamente al cuerpo de una neurona biológica.

$$Neta_i = \sum X_j W_{ij}$$

2.1.3 Función de Activación

Una vez que la entrada neta ha sido calculada, se transforma en el valor de activación, o activación simplemente y una vez hecho esto se puede aplicar la función de salida que es la encargada de transformar el valor de la entrada neta en el valor de salida del nodo [Freeman & Skapura, 1991]. La función de activación F puede ser lineal o no lineal. Existen varios tipos de funciones de activación:

-Función Lineal:

OUT= K(NET) donde K es una constante y NET es una señal

-Función squash, sigmoide o función logística:

$$F(X) = \frac{1}{(1+e^{-NET})}$$

-Función de tangente hiperbólica:

OUT= Tanh (NET)

-Función umbral:

$$OUT = \begin{cases} 1 \text{ si NET}>T \text{ donde T es un valor de umbral} \\ 0 \text{ si no} \end{cases}$$

2.1.4 Clasificacion de las redes neuronales de acuerdo a su complejidad

Los neurones se relacionan entre sí formando redes que pueden llegar a ser tan complejas como el neocognitrón, o tan simples como el perceptrón.

Las RNAs de un nivel (o de una capa). es el modelo más simple según se observa en la Figura 2.2.

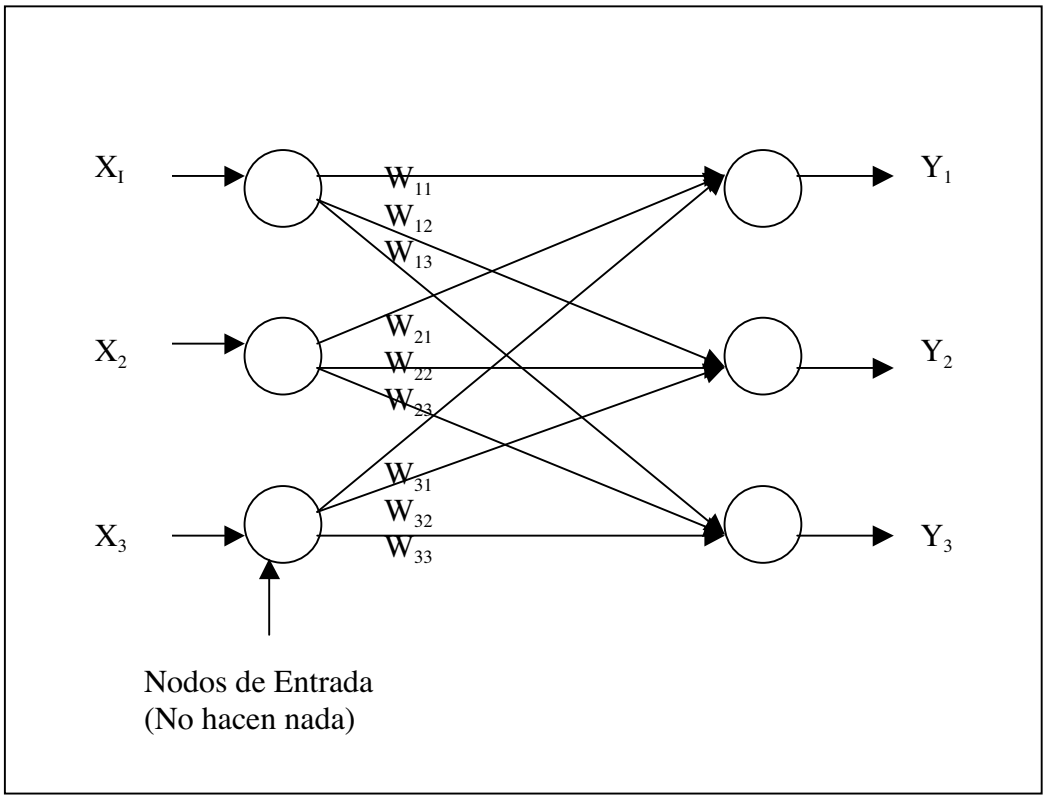

Figura 2.2 Red Neuronal de 1 nivel [Gómez, 1999]

Las RNAs de varios niveles se puden visualizar como lo muestra la figura 2.3. Si existen varios niveles o capas, la función de activación debe ser no lineal, ya que de no ser así una red con varios niveles equivaldría a una red con un nivel.

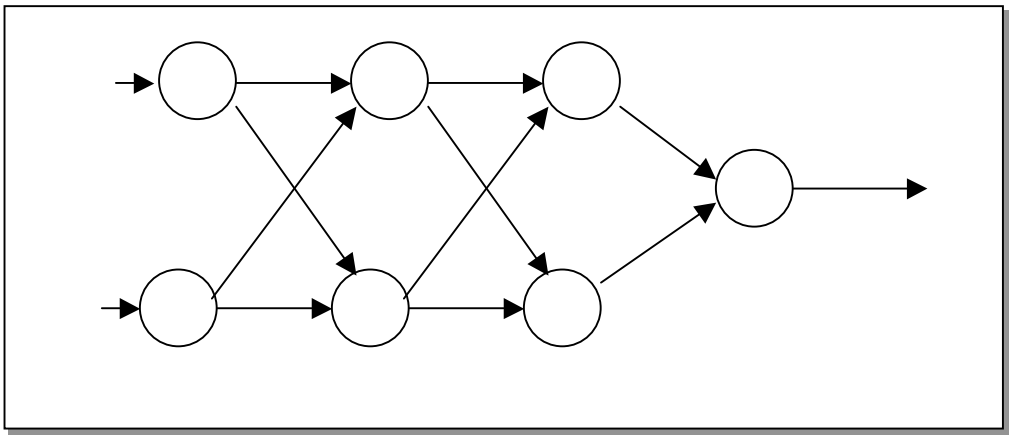

Figura 2.3 Ejemplo de una RNA de varios niveles

2.1.5 Tipos de entrenamiento

El entrenamiento es una de las herramientas que las RNAs nos proporcionan para agilizar el aprendizaje. Este proceso consiste en ir ajustando los pesos W gradualmente hasta que el vector de salida resultante coincida con el vector de salida deseado.

El entrenamiento supervisado parte de un vector de entrada del cual se conoce su vector de salida deseada o al menos una aproximación a él. Al par de vectores representando los valores de entrada y salida deseada se le denomina par de entrenamiento. Este proceso consiste en aplicar el vector de entrada a la red . La diferencia o cambio existente entre el vector de salida y el vector de salida deseada se reduce a través de diversos algoritmos existentes. Se continúa probando diversos vectores de entrada ajustando pesos, hasta que la diferencia con la salida deseada es mínima.

En el entrenamiento no supervisado se desconoce la salida, únicamente se proporciona un vector de entrada. Lo que se busca es generar después de varios vectores de entrada, salidas que sean consistentes. Es decir, que los pesos se vayan ajustando poco a poco a través de el reconocimiento de patrones, regularidades, propiedades estáticas, etc. Así, las entradas similares producirán el mismo tipo de salida. Otra forma de explicar esto es, que este proceso extrae propiedades estadísticas del conjunto de entrenamiento.

Actualmente la mayoría de los algoritmos de entrenamiento se basan en el trabajo desarrollado por D.O. Hebb, quien propuso un algoritmo de entrenamiento sin supervisión donde los pesos W se incrementan si tanto el neurón emisor como el receptor están activados. Este tipo de aprendizaje es el que se adquiere cuando un humano repite una misma tarea. Un ejemplo que se puede mencionar se presenta en los inicios de la medicina, en el que era muy común que un médico buscara constantemente la combinación exacta de ciertas sustancias para lograr que un paciente se curara. La tarea era repetida constantemente hasta que después de varios intentos se descubrían los tipos y cantidades exactos para curar a un paciente en general. Como se puede observar no existía una receta deseada con la cual comparar la salida, sino simplemente se obtuvieron patrones.

2.1.6 Modelo de Retro-Propagación

Este es un algoritmo de aprendizaje que es utilizado para entrenar redes de varios niveles. Fue ideado por E. Rumelhart, G.E. Hinton y R.J. Williamsen 1986 [Gómez, 1999]. Este algoritmo posee una base matemática bastante sólida y que es considerado como una generalización de la regla delta. Esta técnica minimiza el error promedio al cuadrado entre la salida real y la esperada, aplicando el concepto de gradiente descendiente.

El objetivo de Retro-propagación es que los pesos de los niveles escondidos generen una representación interna adecuada al problema a resolverse. Estas características y su porcentaje de éxito lo han convertido en uno de los algoritmos de aprendizaje más populares.

2.2 Segmentación de palabras usando RNAs

Los investigadores han utilizado diferentes técnicas tanto para tareas de segmentación como para tareas de reconocimiento de palabras. Algunos de ellos han utilizado técnicas heurísticas convencionales para la segmentación y el reconocimiento, mientras que otros han usado técnicas heurísticas seguidas de una RNAs basadas en métodos para el reconocimiento de caracteres/palabras. Para escritura impresa o manuscrita, algunos de los resultados más exitosos se han obtenido con el uso de técnicas

que usan componentes combinados. Estas técnicas son léxico-dirigidas y arrojan las mejores imágenes de palabras o caracteres segmentados, las subimágenes de una palabra pueden ser armadas y comparadas para representar una posible palabra dentro de un léxico. Estas técnicas no emplean algoritmos de segmentación muy complicados. Como resultado, el número de puntos de segmentación son muchos. Si se emplean técnicas de segmentación más poderosas es posible reducir el número de puntos de segmentación falsos. Esto puede elevar la velocidad y la eficiencia de los sistemas reduciendo el número de combinaciones primitivas que necesitan ser armadas y comparadas contra el léxico de las palabras.

Recientemente, muchos investigadores han puesto especial atención a las RNAs para ayudar al proceso de segmentación [Blumenstein & Verma, 1999]. Desafortunadamente solo han sido algunos cuantos autores que han detallado sus descubrimientos en letras cursivas. Es por ello que la mayoría de las técnicas de segmentación no están generalmente explicadas en el contexto de un sistema completo, los investigadores tienden a medir el éxito de sus sistemas en base a sus descubrimientos en las fases de reconocimiento de caracteres o palabras.

En resumen podemos inferir que el campo de segmentación de palabras a través de RNAs (o sistemas híbridos que las utilicen) es aún muy amplio y no muy desarrollado debido a que la atención de los investigadores está centrada en el reconocimiento de la palabra, sin tomar la debida importancia a una de las fases más importantes de este proceso: la segmentación.