UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

USO DE REDES NEURONALES ARTIFICIALES PARA EL PRONÓSTICO DE LA INFLACIÓN

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

A C T U A R I A

P R E S E N T A :

ARGELIA YURIKO MELCHOR QUINTO

DIRECTOR DE TESIS:
M. EN A.P. MARÍA DEL PILAR ALONSO REYES
2010

1. Datos del alumno
    Melchor
    Quinto
    Argelia Yuriko
    56 41 99 23
    Universidad Nacional Autónoma de México
    Facultad de Ciencias
    Actuaría
    302045442

2. Datos del Tutor
    M. en A.P.
    María del Pilar
    Alonso
    Reyes

3. Datos del sinodal 1
    Mat.
    Margarita
    Chávez
    Cano

4. Datos del sinodal 2
    Dr.
    Pedro Eduardo
    Miramontes
    Vidal

5. Datos del sinodal 3
    M. en C.
    José Antonio
    Flores
    Díaz

6. Datos del sinodal 4
    M. en I.
    Jorge Luis
    Silva
    Haro

7. Datos del trabajo escrito
    Uso de redes neuronales artificiales para el pronóstico de la inflación
    92 p.
    2010

Pocos son los momentos que he estado sola en mi vida...
Tendría que bendecir al tiempo, al espacio, incluso al azar o al destino, por colocarme en tus brazos madre...en tus palabras padre, y en su corazón hermanas, pero impera en mi la necesidad de mencionar el gran amor que les tengo e ignorar los cómos y porqués de estar en sus vidas y ustedes en la mía, así como agradecerles mis días llenos de su compañía. Y a usted un reconocimiento especial Abu, mi segunda madre, que de recordar mi infancia bajo su cuidado rompo en lágrimas al saberla ahora ya perdida en el tiempo.

INDICE

Introducción

Mejorar y complementar los ejercicios de pronóstico de la inflación es de suma importancia para el diseño e implementación de la política monetaria y, en consecuencia para el logro del objetivo esencial de estabilidad interna y externa de la moneda nacional que tienen los bancos centrales de los países, en particular, el Banco de México en el nuestro. Al mismo tiempo el conocimiento del comportamiento de esta variable puede llegar a ser muy útil para inversionistas y para el público en general.

Existe consenso en que la inflación, en el mediano y largo plazo, tiene un origen estrictamente monetario, razón por la cual es de esperar que los agregados monetarios contengan información útil acerca de la dirección futura de esta variable. Por otra parte, se han realizado diversos estudios que sugieren la existencia de un comportamiento asimétrico entre la política monetaria y el nivel de precios[1]. Así, el supuesto generalmente aceptado de una relación lineal en la modelación de variables económicas podría no justificarse.

Dadas las premisas anteriores, el objetivo de este documento es establecer un modelo para el pronóstico de la inflación, así como mostrar la relación entre el dinero y ésta por medio de un modelo de RNA, todo lo anterior con fundamentos matemáticos que validen su utilización. Esta técnica de modelación, capaz de capturar no linealidades, tiene ventajas en comparación con otros métodos tradicionales de pronóstico. Entre ellas se encuentran: i) aproximar la forma funcional que mejor caracteriza las series de tiempo, ii) superar los métodos de pronóstico tradicionales en horizontes de largo plazo, y iii) ser tan buena como éstos en el pronóstico de corto plazo[2].

[1] Ver Capítulo IV
[2] 5años<Largo plazo, 1año<Mediano plazo< 5 años, corto plazo< 1 año.

El presente trabajo esta dividido en cuatro capítulos. El primero de ellos abarca de forma breve el análisis de series de tiempo en forma clásica y bajo qué condiciones éstas se consideran procesos estocásticos, lo anterior con el fin de ubicar a las RNA dentro de dichos modelos y de introducir al lector en el estudio de las series de tiempo.

El segundo capítulo describe el comportamiento y características de las RNA y se centra en el funcionamiento del Perceptron Multicapa, tipo de estructura de red utilizada para el pronóstico de la inflación en el escrito.

En el capítulo tres se define la variable a pronosticar así como su medición, características, repercusiones económicas y variables causales y correlacionadas.

En el último capítulo se procede al pronóstico de inflación con el uso de RNA.

Capítulo I
Series de Tiempo

1.1. Definición

Una serie de tiempo puede definirse como una sucesión de observaciones correspondientes a una variable en distintos momentos de tiempo realizadas a intervalos regulares y de duración constante. Por lo anterior las series pueden tener una periodicidad anual, semestral, mensual, etc. según los periodos de tiempo en que sean recogidos los datos que la componen.

En todas estas variables los datos pueden referirse, o bien a un intervalo de tiempo, en cuyo caso la variable recibe la denominación de *flujos* o, por lo contrario, la observación puede haberse recogido por instante concreto de tiempo, hablando entonces de *stock*.

1.2. Análisis de series de tiempo

La sistematización de las diferentes metodologías empleadas en el tratamiento de una serie de tiempo puede hacerse con base a criterios diversos, que dan lugar a los siguientes enfoques:

1. *Determinista y estocástico*. En el primero, la variable observada se supone que presenta un patrón de comportamiento fijo o determinista. Esto significa que las irregularidades de la serie se contemplan como una desviación respecto a una pauta de comportamiento sistemática. Este enfoque es el que está presente en el análisis clásico o de descomposición de series de tiempo. Frente al enfoque determinista surge alrededor de los años veinte una concepción distinta de percibir la generación de una serie a través de

la teoría de los procesos estocásticos, donde la metodología de Box-Jenkins es la que más difusión ha tenido.

2. *Univariante y multivariante*. El enfoque univariante trata de explicar la trayectoria de una variable a través de la información contenida en los datos históricos de su correspondiente serie, es decir, intenta capturar el comportamiento sistemático que muestre el pasado de la misma y con base a ello realizar predicciones. En el multivariante, la metodología empleada es la del análisis causal, a través de la incorporación de factores externos como causa de la explicación del comportamiento de una determinada variable.

3. *Dominio temporal y frecuencial*. El tratamiento de una serie en un dominio temporal tiene como finalidad la elaboración de modelos dependientes del tiempo. En un análisis frecuencial o espectral, la variable independiente pasaría a ser la frecuencial, persiguiéndose con ello la detección de las fluctuaciones periódicas de una serie, investigando la estructura interna del proceso estocástico que la genera.

1.2.1. Objetivo del análisis de series de tiempo

Con el análisis de series de tiempo se pretende extraer las regularidades, que se observan en el comportamiento pasado de la variable, obtener el mecanismo que la genera, para que con base en ello, tener un mejor conocimiento de la misma en el tiempo y, bajo el supuesto de que las condiciones estructurales que conforman el fenómeno objeto de estudio permanecen constantes, predecir el comportamiento futuro reduciendo, de esta forma, la incertidumbre en la toma de decisiones.

El análisis de series de tiempo puede, por tanto, hacerse con dos propósitos:

1. Describir la evolución que la serie ha tenido en el pasado.
2. Predecir sus valores respecto a un futuro más o menos cercano.

1.2.2. Etapas que se necesitan cubrir en el análisis de series de tiempo

1. Notas sobre el tratamiento de los datos que generan una serie de tiempo.

    a. Conocimiento de la relación existente entre la estadística y la variable que se quiere analizar.
    b. La serie de tiempo representa una medición de un fenómeno, que no es indiferente a los procedimientos estadísticos empleados en su confección. La forma en la que ha sido elaborada la estadística y que cambios metodológicos ha sufrido a lo largo del tiempo serán cuestiones que deben ser tomadas en cuenta para trabajar con datos homogéneos. También se tiene que poner atención en la manera en la que se han obtenido los datos: muestreo, registro, censo, etc.
    c. La interpretación de algunos resultados pueden variar al considerar al tiempo como variable en los modelos que se formulan teóricamente, pues se considera continuo, y en las aplicaciones empíricas se trata de forma discreta debido a la naturaleza de los datos.
    d. Es necesario homogeneizar la serie, eliminado los *efectos de calendario* propiciados por el número de días distinto entre los meses. Lo que suele hacerse en la práctica, es multiplicar por un factor de ajuste para que todos los meses tengan el mismo número de días (series estacionales o desestacionalizadas)
    e. Se define la *autocorrelación* como la relación de dependencia lineal que tiene una variable consigo misma, la cual permite concluir que la dependencia temporal que se observa en las series de tiempo requiere un tratamiento de datos distinto a las que no la poseen y es

un factor que de no ser tomado en cuenta proporcionará conclusiones no del todo acertadas.

2. Representación gráfica de la serie

La inspección del gráfico correspondiente a los datos de la variable permite tener una idea general del comportamiento de la serie, detectando, como primera observación, posibles tendencias crecientes o decrecientes, influencias estacionales, algún valor muy superior o inferior a lo normal, etc. La exploración gráfica también facilitará información acerca de los métodos más adecuados que se aplicarán en cada caso.

3. Modelización

Esta etapa consiste en elaborar una representación simplificada de las características más importantes que contiene la serie relacionadas con su evolución en el tiempo. Es importante conocer tanto el funcionamiento de los métodos empleados como las condiciones que se necesitan cumplir para su correcta aplicación.

4. Validación del modelo

Superada la fase de modelización, es necesario conocer si el modelo que representa la serie tiene validez tanto para describir su historia como hacer predicciones. Existen numerosos procedimientos para la valoración de los modelos y para evaluar la capacidad predictiva de los mismos. La validez del modelo dependerá de tres factores:

a. Que los errores que presente el modelo no sean muy significativos.
b. Que exista una cierta estabilidad en la estructura del fenómeno, de tal forma que el comportamiento pasado de la variable permanezca en el futuro.
c. Que los datos sean homogéneos en el tiempo.

5. Predicción

Si se llega a un modelo que represente la evolución temporal de la serie, que funcione acertadamente y que haya superado las pruebas de validez a las que se haya sometido, es posible efectuar predicciones.

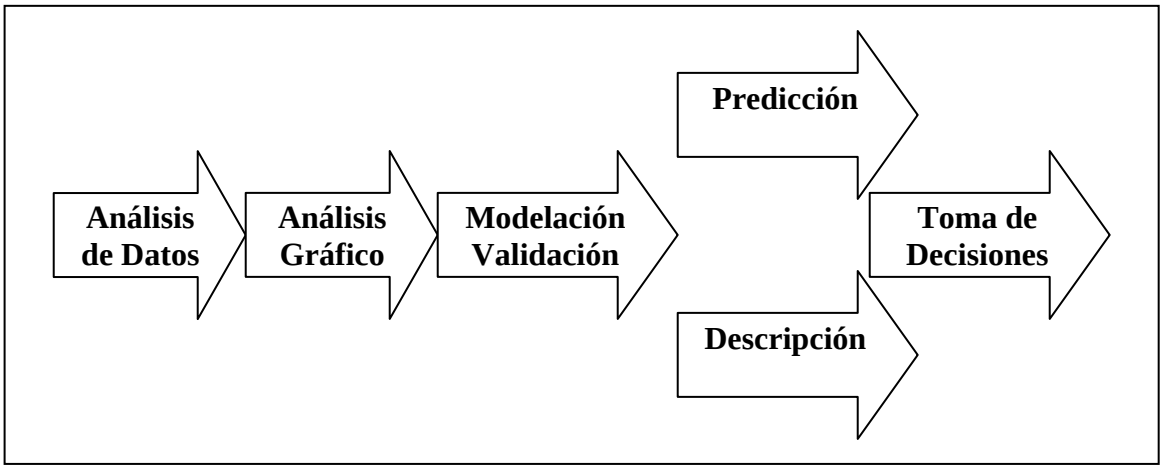


Figura 1.1. *Etapas necesarias para el estudio de una serie de tiempo.*

1.2.3. Análisis clásico de series de tiempo

Desde un marco de referencia clásico, una serie se concibe integrada por movimientos, a corto, medio y largo plazo, más o menos superpuestos. Se tratará de describir las pautas de regularidad que sigue cada unos de estos movimientos, con la finalidad de reproducir el comportamiento de dicha serie.

Desde una perspectiva teórica y con el objetivo de comprender mejor la evolución de un determinado fenómeno, el enfoque clásico o de descomposición de series de tiempo considera que el comportamiento de una variable en el tiempo es el resultado de la integración de cuatro componentes fundamentales: tendencia, ciclo, componente estacional y componente irregular.

Tendencia. Se considera tendencia al movimiento suave y regular de la serie a largo plazo. Es una componente que reviste gran interés ya que refleja la dirección del movimiento de una determinada variable. De esta manera puede detectarse si,

a largo plazo, la serie adopta una marcha persistente, ya sea de crecimiento, decrecimiento o estabilidad, aunque para observar esto es necesario tener un horizonte temporal más amplio. La predicción de esta componente suele ser en muchos casos el objetivo del análisis de series de tiempo a través de los modelos de ajuste de tendencia, donde se supone que la serie carece de variaciones estacionales y cíclicas.

Ciclo. En las series, sobre todo en las económicas, la tendencia puede disgregarse a su vez en un factor de tendencia pura, que representa la evolución suave y continua a largo plazo, y un factor oscilante, caracterizado por movimientos recurrentes en torno a la tendencia y al que se denomina “ciclo económico” el cual se distingue por una serie de movimientos ascendentes y descendentes separados por puntos de inflexión que en la terminología económica corresponden a las denominadas fases de recuperación, prosperidad, recesión y depresión.

Estacionalidad. Se puede definir como los movimientos regulares de la serie que tiene una periodicidad menor a un año. Recoge, por tanto, las oscilaciones que se realizan con esta periodicidad. Existen muchos casos donde se observa claramente esta componente y la mayor parte de las veces obedece a factores institucionales o climatológicos.

Componente irregular. Se caracteriza porque no responde a un comportamiento sistemático o regular su finalidad es distinguir aquellas irregularidades cuyas causas se pueden identificar (*factor errático*) y aquellas atribuibles al azar (*factor aleatorio*).

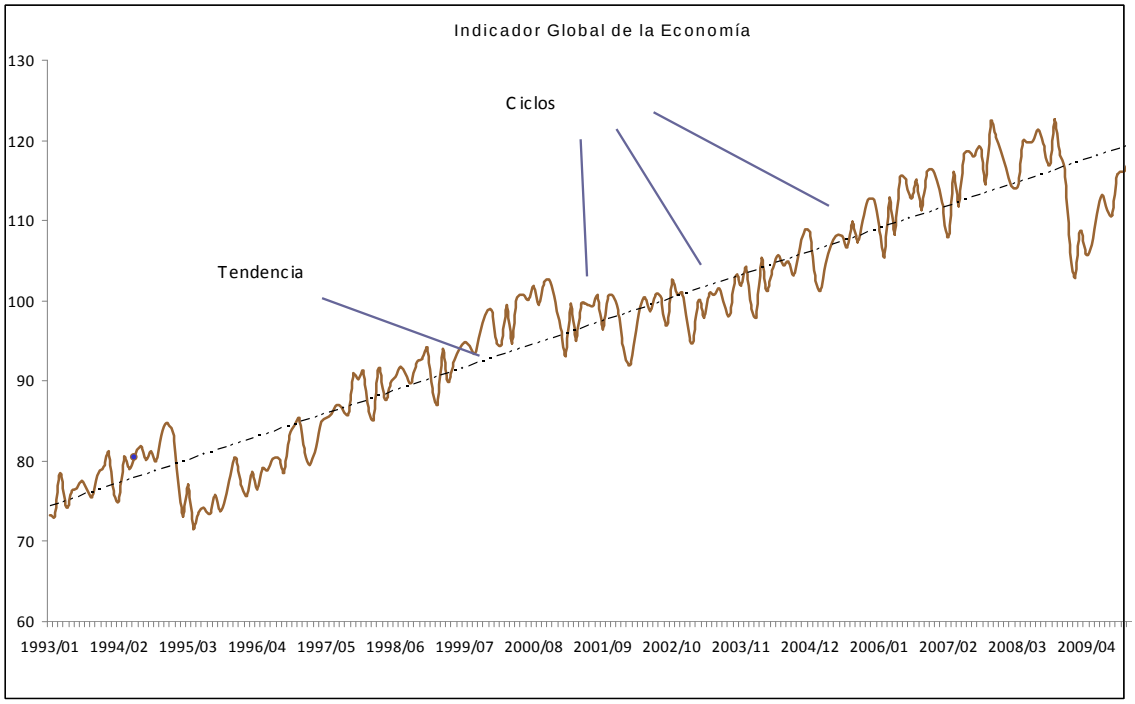


Figura 1.2. Elementos del Análisis Clásico de Series de Tiempo.

1.3. Análisis estocástico de series de tiempo

Para describir que es una serie de tiempo en el contexto de los procesos estocásticos es necesario definir este tipo de proceso: es una familia de variables aleatorias asociadas a un conjunto índice de números reales, de tal forma que a cada elemento del conjunto le corresponda una y sólo una de la variables aleatorias, es decir, $\{X(\tau);\tau \in T\}$, en donde T es el conjunto índice y $\{X(\tau)\}$ es la variable aleatoria correspondiente al elemento τ de T. Si T es un intervalo de números reales, ya sea cerrado o abierto, se dirá que el proceso estocástico es continuo, y si T es un conjunto finito o infinito pero numerable se dirá que es discreto. El hecho de que el proceso estocástico sea continuo o discreto no indica nada acerca de la naturaleza de las variables aleatorias involucradas, ya que éstas son continuas o discretas.

Con base en lo anterior una serie de tiempo es la sucesión de observaciones generadas por un proceso estocástico, cuyo conjunto índice se toma con respecto

al tiempo. Por tanto la inferencia que se realice será acerca de las características del proceso generador de la serie observada. Además así como existen procesos estocásticos discretos y continuos existen series de tiempo discretas y continuas. En particular si las observaciones de una serie de tiempo discreta se toman en los momentos $\{\tau_1, \tau_2, ..., \tau_N\}$, el proceso estocástico respectivo se denotará por $\{X(\tau_1), X(\tau_2), ..., X(\tau_N)\}$.

En adelante se considerarán exclusivamente series de tiempo discretas, con la característica adicional de que las observaciones se hagan en intervalos de tiempo iguales. Asimismo, en general no se hará distinción entre una variable aleatoria X y su valor observado, que se denotará también por X. De esta manera cuando se tengan N valores sucesivos de una serie de tiempo se escribirá como $X_1, X_2, ..., X_N$.

Es importante notar que una serie de tiempo observada no es más que una realización de un proceso estocástico, lo cual significa que bien pudo haberse observado otra realización del mismo proceso, pero cuyo comportamiento fuese distinto del que se observó en realidad. Con lo anterior se pretende subrayar el elemento probabilístico presente en una serie de tiempo, ese mismo elemento será el que conduzca a tener en cuenta la función de densidad conjunta de las variables aleatorias que constituyen el proceso estocástico.

El comportamiento de N variables aleatorias puede describirse con base a su función de densidad conjunta $f(X_1, X_2, ..., X_N)$. En casi todo el análisis estadístico, excepto en el de series de tiempo, se supone que las observaciones que se tienen provienen de variables aleatorias independientes, de tal forma que con el conocimiento de las funciones de densidad individuales, es posible obtener fácilmente la función de densidad conjunta. En contraste, en el caso de las series de tiempo se supone que existe toda una estructura de correlación entre las observaciones; por consiguiente, no es posible obtener la función de densidad

conjunta de manera tan directa y debe utilizarse alguna otra forma para caracterizar las variables aleatorias que intervienen.

1.3.1. Procesos estacionarios

Se dice que una serie de tiempo es estrictamente estacionaria si la distribución conjunta de $X(t_1),...,X(t_k)$ es la misma a la distribución conjunta de $X(t_1+\tau),...,X(t_k+\tau)$ para toda $t_1,...,t_k,\tau$ y cualquier k. En otras palabras, cambiar el origen en un monto τ no afecta la distribución conjunta, por lo tanto depende sólo de los intervalos entre $t_1,...,t_k$. A τ se le da el nombre de *retrazo*.

Para un proceso estrictamente estacionario, dado que la función de distribución es la misma para toda t, la media $\mu_t = \mu$ es constante, tal que $E(|X_t|) < \infty$.

Un proceso estacionario es de segundo orden (o débilmente estacionario) si su media es constante y su coeficiente de autocorrelación depende sólo del retrazo, es decir:

$$E[X(t)] = \mu \tag{1.1}$$

$$Cov[X(t), X(t+\tau)] = \gamma(\tau) \tag{1.2}$$

$\tau = 0$ implica que la varianza es al igual que la media, constante. La definición anterior también implica que tanto varianza y media son finitas.

1.3.2. Procesos aleatorios puros

Un proceso discreto es llamado un proceso aleatorio puro si consiste en una secuencia de variables aleatorias $\{Z_t\}$, las cuales son mutuamente independientes e idénticamente distribuidas. Generalmente se asumen con distribución normal de media cero y varianza σ_Z^2. Por definición se sigue que el proceso tiene media y varianza constantes.

Los procesos aleatorios puros son usados en muchas situaciones, particularmente para construir procesos más complicados que serán mencionados posteriormente. Algunos autores prefieren asumir que las $Z_t´s$ son mutuamente no correlacionadas en vez de independientes. Lo anterior es adecuado para los modelos lineales y normales, pero se hace necesario pedir la independencia cuando se trabaja con modelos no lineales. Los procesos aleatorios puros comúnmente son llamados ruido blanco.

1.3.3. Modelos lineales

Los modelos lineales para los procesos estocásticos se basan en la idea de que una serie de tiempo, cuyos valores sucesivos pueden ser altamente dependientes, puede ser generada a partir de una serie de choques aleatorios independientes $\{a_t\}$ resultado de realizaciones independientes de una variable aleatoria cuya media es constante (por lo general se considera igual a cero) y cuya varianza es σ_a^2. A esta sucesión de variables aleatorias se le conoce como ruido blanco. La idea previa lleva a representar al proceso $\{Z_t\}$ en función de $\{a_t\}$ mediante la relación lineal

$$Z_t = \mu + a_t - +\psi_1 a_{t-1} - \psi_2 a_2 - ... \quad (1.3)$$

en donde μ es un parámetro que determina el nivel (no necesariamente la media) del proceso y $\psi(B)$ es un polinomio de retrazo denotado por

$$\psi(B) = 1 - \psi_1 B - \psi_2 B^2 - ... \quad (1.4)$$

que convierte el proceso $\{a_t\}$ en el proceso $\{Z_t\}$ mediante un filtro lineal basado en el operador lineal $\psi(B)$.

1.3.3.1. Procesos de promedios móviles (MA (q))

Supóngase que $\{Z_t\}$ es un proceso aleatorio puro. El proceso $\{X_t\}$ se dice que es un proceso de promedios móviles de orden q si:

$$X_t = \beta_0 Z_t + \beta_1 Z_{t-1} + ... + \beta_q Z_{t-q} \quad (1.5)$$

donde $\{\beta_i\}$ son constantes y las Z normalmente se escalan por lo que $\beta_0 = 1$. No se requieren restricciones específicas para que el proceso MA sea estacionario, pero generalmente es deseable imponer ciertas condiciones para asegurar que satisfaga la condición de invertibilidad.

Se dice que un proceso $\{X_t\}$ es invertible si una perturbación aleatoria en el tiempo t, algunas veces llamada *innovación*, puede ser expresada, como una suma convergente del valor presente y valores pasados de X_t

$$Z_t = \sum_{j=0}^{\infty} \pi_j X_{t-j} \quad (1.6)$$

donde $\sum |\pi_j| < \infty$. Lo anterior significa que el proceso puede reescribirse como un modelo autorregresivo, posiblemente de orden infinito, cuyos coeficientes forman parte de la suma convergente.

La condición de invertibilidad para un proceso MA de cualquier orden se expresa de manera más simplificada con la ayuda del *operador de retrazo* denotado por B definido de la siguiente manera

$$B^j X_t = X_{t-j} \qquad \forall j \quad (1.7)$$

por lo que la ecuación (1.5) se puede reescribir

$$X_t = (\beta_0 + \beta_1 B + ... + \beta_q B^q) Z_t = \theta(B) Z_t \quad (1.8)$$

donde $\theta(B)$ es un polinomio de orden q en B. Es posible probar que un proceso MA(q) es invertible si todas las raíces de la ecuación

$$\theta(B) = \beta_0 + \beta_1 B + ... + \beta_q B^q = 0 \quad (1.9)$$

se encuentran fuera del círculo unitario, donde se considera a B como una variable compleja y no como un operador. Esto significa que las raíces, que pueden ser o no complejas, tienen módulo mayor a la unidad.

Los procesos MA han sido utilizados en diferentes campos, particularmente en la econometría. Por ejemplo, varios indicadores económicos son afectados por una gran variedad de eventos aleatorios que pueden no tener una influencia inmediata sobre dichos indicadores pero pueden afectarlos en periodos subsecuentes, lo cual hace que un proceso MA sea apropiado para su modelación.

1.3.3.2. Procesos autorregresivos (AR(p))

Supóngase que $\{Z_t\}$ es un proceso aleatorio puro. El proceso $\{X_t\}$ se dice que es un proceso autorregresivo de orden p si:

$$X_t = \alpha_1 X_{t-1} + ... + \alpha_p X_{t-p} + Z_t \quad (1.10)$$

El modelo anterior es parecido a un modelo de regresión, sin embargo, en este caso X_t está representada por valores de ella misma en tiempos pasados más que por variables predictoras separadas, lo que da origen al prefijo *auto*.

Se observa que para el caso de primer orden (*p=1*), $X_t = \alpha_1 X_{t-1} + Z_t$, se puede realizar la sustitución de manera sucesiva

$$X_t = \alpha(\alpha X_{t-2} + Z_{t-1}) + Z_t \quad (1.11)$$

$$X_t = \alpha^2(\alpha X_{t-3} + Z_{t-2}) + \alpha Z_{t-1} + Z_t \quad (1.12)$$

y eventualmente se encuentra que X_t puede ser expresada como un proceso MA de orden infinito de la forma

$$X_t = Z_t + \alpha Z_{t-1} + \alpha^2 Z_{t-2} + ... \quad (1.13)$$

que converge con $-1 < \alpha < 1$. El que un proceso AR pueda ser escrito en la forma de un proceso MA señala que existe una dualidad entre un AR y un MA.
Un proceso AR es estacionario si $|\alpha|$ es estrictamente menor a la unidad.

1.3.3.3. Modelos ARMA (p , q)

Con la combinación de un proceso AR con p parámetros y un proceso MA de orden q se crea un modelo llamado proceso ARMA (p , q) dado por

$$X_t = \alpha_1 X_{t-1} + ... + \alpha_p X_{t-p} + Z_t + \beta_1 Z_{t-1} + ... + \beta_q Z_{t-q} \quad (1.14)$$

usando el operador de retrazo en la ecuación anterior puede reescribirse como sigue

$$\phi(B) X_t = \theta(B) Z_t \quad (1.15)$$

donde $\phi(B)$ y $\theta(B)$ son polinomios de orden p y q respectivamente.

Las condiciones para que un proceso ARMA sea estacionario e invertible es el mismo que para los procesos MA y AR puros.

La importancia de un modelo ARMA radica en el hecho de que las series de tiempo estacionarias pueden ser modeladas adecuadamente por un proceso de este tipo con un número menor de parámetros que los que se utilizarían cuando se modela con procesos AR y MA puros, es decir, se cumple el principio de *parsimonia*, el cual señala que es posible obtener una adecuada representación de los datos con el menor número de parámetros posibles.

En algunas ocasiones es de gran utilidad expresar un modelo ARMA como un proceso MA puro en la forma

$$X_t = \varphi(B) Z_t \quad (1.16)$$

donde $\varphi(B) = \theta(B)/\phi(B)$. Alternativamente, también puede expresarse un modelo ARMA como un proceso AR puro

$$\pi(B) X_t = Z_t \quad (1.17)$$

donde $\pi(B) = \phi(B)/\theta(B)$.

1.3.3.4.Modelos ARMA integrados o ARIMA (p , d , q)

En la práctica muchas series son no estacionarias. Para poder utilizar los modelos referidos anteriormente se hace necesario remover las fuentes no estacionarias de variación, con este fin es recomendable obtener la serie de diferencias de la

original que posee en comportamiento más suave y regular, reemplazando X_t por $\nabla^d X_t$ en la ecuación (1.14). Dicho modelo es llamado “modelo integrado”.

$$W_t = \nabla^d X_t = (1-B)^d X_t \tag{1.18}$$

por lo que el modelo ARIMA tiene la siguiente forma

$$W_t = \alpha_1 W_{t-1} + ... + \alpha_p W_{t-p} + Z_t + \beta_1 Z_{t-1} + ... + \beta_q Z_{t-q} \tag{1.19}$$

$$\phi(B) W_t = \theta(B) Z_t \tag{1.20}$$

$$\phi(B)(1-B)^d X_t = \theta(B) Z_t \tag{1.21}$$

1.3.4. Modelos no lineales

Muchos modelos ARMA no lineales pueden ser vistos como casos especiales de la siguiente forma general

$$\begin{aligned} &z_t - \Phi_1(Y_{t-1}) Z_{t-1} - ... - \Phi_p(Y_{t-1}) Z_{t-p} = \\ &\theta_0(Y_{t-1}) + a_{t-1} - \theta_1(Y_{t-1}) a_{t-1} - ... - \theta_q(Y_{t-1}) a_{t-q} \end{aligned} \tag{1.22}$$

donde

$$Y_{t-1} = (z_{t-1,...,} z_{t-p}, a_{t-1,...,} a_{t-q})' \tag{1.23}$$

y $\Phi_i(Y_{t-1})$, $\theta_i(Y_{t-1})$ son vectores del vector Y_{t-1} en el tiempo $t-1$. Para estos casos específicos se consideran las siguientes variantes.

1.3.4.1. Modelos bilineales

Sean Φ_i constantes y el conjunto $\theta_j(Y_{t-1}) = b_j + \sum_{i=1}^{k} bij z_{t-i}$ se tiene el siguiente modelo:

$$z_t - \Phi_1 Z_{t-1} - \ldots - \Phi_p Z_{t-p} =$$
$$\theta_0(Y_{t-1}) + a_t - \sum_{j=1}^{q} bja_{t-j} - \sum_{i=1}^{k}\sum_{j=1}^{q} bijz_{t-i}a_{t-j} \quad (1.24)$$

1.3.4.2. Modelos TAR, modelos autorregresivos con umbrales
(Threshold autoregressive Model)

Sea $\theta_i = 0, i \geq 1$, d rezago de tiempo entero y c una constante el umbral

$$\Phi_i(Y_{t-1}) = \begin{Bmatrix} \Phi_i^{(1)} \text{ si } z_{t-d} \leq c \\ \Phi_i^{(2)} \text{ si } z_{t-d} > c \end{Bmatrix}$$
$$\theta_0(Y_{t-1}) = \begin{Bmatrix} \theta_0^{(1)} \text{ si } z_{t-d} \leq c \\ \theta_0^{(2)} \text{ si } z_{t-d} > c \end{Bmatrix} \quad (1.25)$$

se tiene el modelo

$$z_t = \begin{Bmatrix} \theta_0^{(1)} + \sum_{i=1}^{p} \Phi_i^{(1)} z_{t-i} + a_t^{(1)} \text{ si } z_{t-d} \leq c \\ \theta_0^{(2)} + \sum_{i=1}^{p} \Phi_i^{(2)} z_{t-i} + a_t^{(2)} \text{ si } z_{t-d} > c \end{Bmatrix} \quad (1.26)$$

donde $\{a_t^{(1)}\}$ y $\{a_t^{(2)}\}$ son procesos de ruido blanco y el valor c es llamado el parámetro umbral y d es el parámetro de rezago. La definición anterior se puede extender a un modelo que contemple l umbrales de la forma siguiente:

$$z_t = \theta_0^{(j)} + \sum_{i=1}^{p} \Phi_i^{(j)} z_{t-i} + a_t^{(j)} \text{ si } c_{j-1} \leq z_{t-d} \leq c_j \quad j = 1, \ldots l \quad (1.27)$$

donde $c_1 < c_2 < \ldots < c_{t-1}$

1.3.4.3. Modelos autoregresivos (Nonlinear autoregressive models- NAR)

Estos modelos se caracterizan porque recogen el comportamiento temporal de la serie expresando el valor de la misma en el instante $t+1$ como una función no lineal de $r+1$ valores de la serie en instantes anteriores de tiempo, es decir:

$$x(t+1) = F(x(t), x(t-1), x(t-2), ..., x(t-r)) + \varepsilon(t) \quad (1.28)$$

donde t es la variable discreta tiempo; $\varepsilon(t)$ es un error residual que se asume ruido blanco y F es una función no lineal desconocida y que, por tanto, debe ser estimada o aproximada a partir de un conjunto de datos observados de la serie de tiempo.

Por tanto, la construcción de modelos NAR involucra la determinación de la función, a partir de muestras disponibles, mediante técnicas de aproximación, entre las cuales se encuentran las redes neuronales artificiales.

Capítulo II
Redes neuronales artificiales

Desde sus principios, la inteligencia artificial ha estado enfocada en mejorar el amplio campo de la ciencia computacional y ha contribuido considerablemente a la investigación en varias áreas científicas y técnicas.

McCulloch y Pitts (1943) y otros autores como Wiener (1985) y Von Neuman (1958), en sus estudios sobre cibernética y su teoría sobre autómatas[3], fueron los primeros en abordar la integración del proceso biológico con métodos de ingeniería. McCulloch y Pitts propusieron el modelo neuronal que ahora lleva su nombre: un mecanismo binario con dos estados y un umbral fijo que recibe estímulos (sinapsis), todos con el mismo valor y ajenos de acciones externas. Ellos simplificaron la estructura y funcionamiento de las neuronas, considerando modelos con *m* datos de entrada (*inputs, ejemplos o patrones*) y sólo uno de salida (*output*), y dos posibles estados: activos o inactivos. En su estado inicial, una red artificial neuronal fue una colección de neuronas de McCulloch y Pitts, todas con la misma escala, y en las cuales los *outputs* de algunas neuronas estaban conectados a *inputs* de otras.

Todas las formalizaciones matemáticas sobre las redes neuronales artificiales que fueron elaboradas después de los estudios referidos han usado sistemas biológicos como un punto de partida para el estudio de redes biológicas neuronales, sin pretender modelos exactos.

[3] Un autómata es una máquina que imita la figura y los movimientos de un ser animado

2.2. Modelo de una neurona biológica

En el área de redes neuronales biológicas los investigadores han analizado varios modelos para explicar cómo funcionan las células del cerebro humano. Como se observa en la figura (2.1) las células contienen un núcleo y una membrana externa eléctrica. Cada neurona tiene un nivel de activación, con rangos entre un máximo y un mínimo.

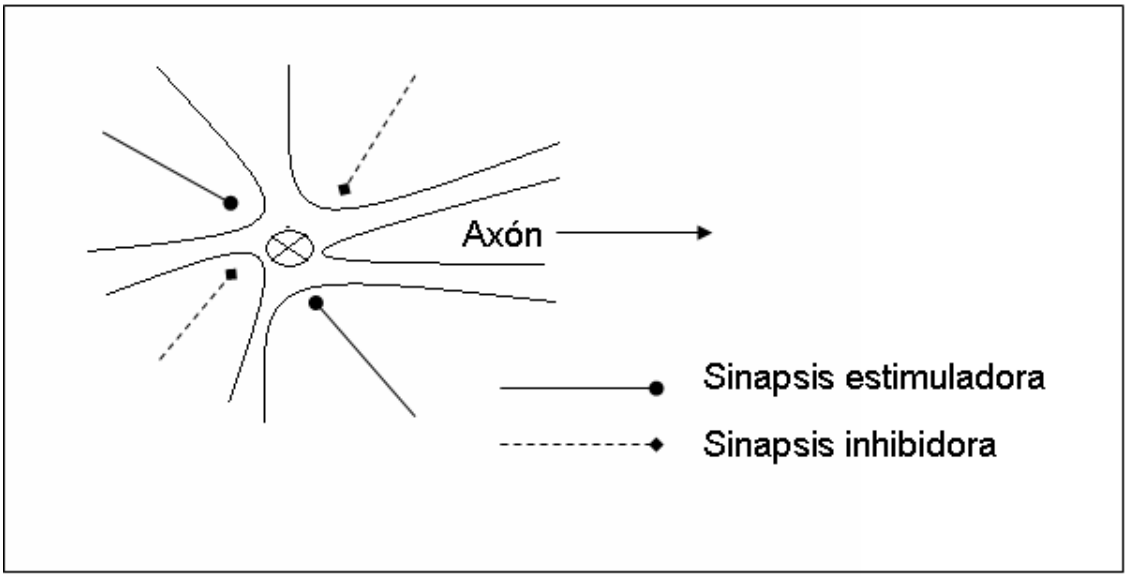


Figura 2.1.*Modelo de una red neuronal biológica.*

El aumento o disminución de la activación de una neurona producto de la relación con otra se da por medio de la sinapsis o espacio de comunicación. Esta sinapsis traslada el nivel de activación desde una célula emisora a una receptora. Si la sinapsis es del tipo estimuladora, el nivel de activación proveniente de la neurona emisora incrementa la de la célula receptora, caso contrario cuando es del tipo inhibidora. La sinapsis no solo difiere por el hecho de ser estimuladora o inhibidora sino también por la intensidad del estímulo. El resultado de cada neurona es transferido por el axón para influir sobre otras.

Este simple modelo neuronal biológico subyace en la mayoría de las aplicaciones de redes neuronales artificiales en la actualidad. Aunque el modelo es únicamente

una aproximación muy gruesa de la realidad, sus exitosas aplicaciones han probado los beneficios de las redes neuronales.

2.3. Redes neuronales artificiales (RNA)

El comportamiento del cerebro tiene varias características altamente deseables para los sistemas de procesamiento digital: es robusto y tolerante a las fallas, las neuronas mueren cada día sin afectar el funcionamiento global del cerebro; puede también manejar información difusa (inconsistente o con "ruido"), es efectivo en tiempo (trabaja de manera paralela), es pequeño, compacto y consume poca energía.

Las RNA se parecen a las redes biológicas neuronales en que no requieren de la programación de tareas pues generaliza y aprende de la experiencia. Los modelos actuales de RNA están compuestos de un conjunto simple de elementos de procesamiento (PE), que emulan de las neuronas biológicas o nodos y de un cierto número de conexiones entre ellas que no ejecutan instrucciones sino que responden en paralelo a los *inputs* presentados, y pueden funcionar correctamente siempre que un PE o una conexión detengan su funcionamiento, o la información tenga un cierto nivel de ruido. Por lo que es un sistema tolerante a ciertas fallas y ruidos, y puede aprender en un proceso de entrenamiento y modificar los valores asociados a las conexiones de los PE para ajustar el *output* ofrecido como respuesta de los *inputs*. Este resultado no se almacena en la memoria del proceso, este es el estado de la red en el cual un balance es alcanzado. La capacidad de una RNA no reside en sus instrucciones pero si en su topología, es decir, de las posiciones de los PE y de las conexiones entre ellos, y de los valores de las conexiones (pesos) entre los PE, y las funciones que definen sus elementos y los mecanismos de aprendizaje.

Las RNA ofrecen una alternativa a la computación clásica para problemas del mundo real, que usan un conocimiento natural (que pueden ser inciertos,

imprecisos, inconsistentes e incompletos) y para los cuales el desarrollo de un programa convencional que cubra todas las posibilidades y eventualidades es impensable o en el mejor de los casos muy laborioso y caro.

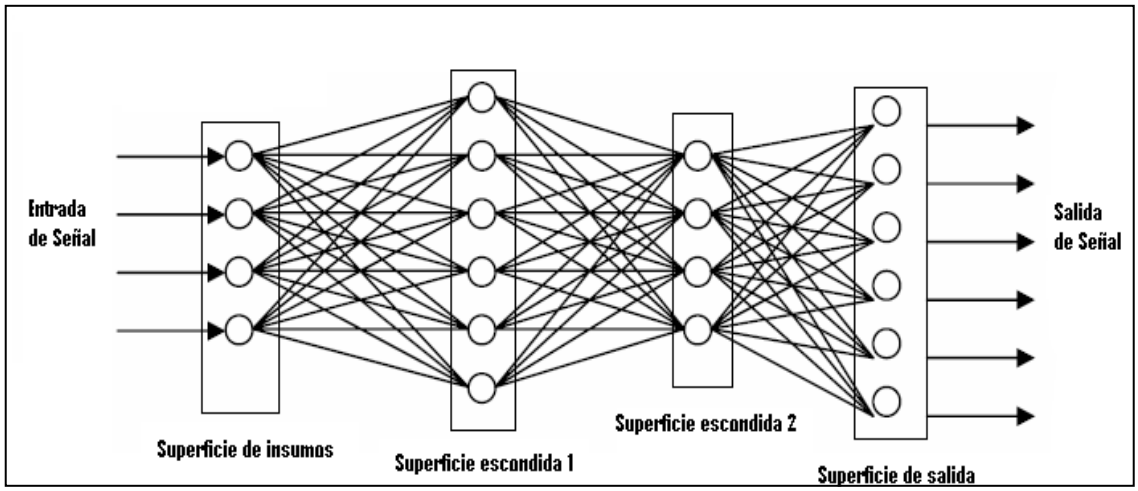


Figura 2.2.*Representación gráfica de una red neuronal artificial.*

2.4. Primeros modelos de redes neuronales artificiales

2.4.1. Perceptron

Este modelo se concibió como un sistema capaz de realizar tareas de clasificación de forma automática. La idea era disponer de un sistema que, a partir de un conjunto de ejemplos de clases diferentes, fuera capaz de determinar las ecuaciones de las superficies que hacían la frontera de dichas clases. La información sobre la que se basaba estaba constituida por los ejemplos existentes de los distintos grupos. Son dichos ejemplos los que aportan la información necesaria para que el sistema construya las superficies discriminantes, y además actuará como un clasificador para ejemplos nuevos desconocidos. El sistema, al final del proceso, era capaz de determinar, para cualquier ejemplo nuevo, a qué clase pertenecía.

La arquitectura de la red es muy simple. Se trata de una estructura monocapa, en la que hay un conjunto de células de entrada, tantas como sea necesario, según los términos del problema, y una o varias células de salida. Cada una de las células de entrada tiene conexiones con todas las células de salida y son estas conexiones las que determinan las superficies de discriminación del sistema.

En este modelo la salida de la red se obtiene de la siguiente forma:

1. Se calcula la activación de la neurona de salida mediante la suma ponderada por los pesos de todas las entradas:

$$y' = \sum_{i=1}^{n} w_i x_i \qquad (2.1)$$

2. La salida definitiva se produce al aplicar una función de salida al nivel de activación de la célula. En un Perceptron ésta es una función escalón que depende del umbral

$$y = F(y', \theta) \qquad (2.2)$$

$$F(s,\theta) = \begin{cases} +1 & \text{si } s > \theta \\ -1 & \text{en otro caso} \end{cases} \qquad (2.3)$$

Simplemente pasando el término θ al otro lado de la ecuación, la salida se puede escribir en una sola ecuación:

$$y = F(\sum_{i=1}^{n} w_i x_i + \theta) \qquad (2.4)$$

donde F ya no depende de ningún tipo de parámetro :

$$F(s,\theta)=\begin{cases}+1 & \text{si } s > 0\\ -1 & \text{en otro caso}\end{cases} \qquad (2.5)$$

Esta ecuación equivale a introducir artificialmente en la salida un nuevo peso θ que no está conectado a ninguna entrada, sino a una ficticia con un valor constante de -1.

La función de salida F es binaria y de gran utilidad en este modelo ya que al tratarse de un discriminador de clases, una salida binaria puede ser fácilmente traducible a una clasificación en dos categorías de la siguiente forma:

- Si la red produce salida 1, la entrada pertenece a la categoría A.

- Si la red produce salida -1, la entrada pertenece a la categoría B.

Para el caso de dos dimensiones la ecuación *(2.4)* se transforma en la ecuación de una recta:

$$w_1 x_1 + w_2 x_2 + \theta = 0 \qquad (2.6)$$

Para más dimensiones esta ecuación es un hiperplano.

El proceso de aprendizaje en esta estructura se lleva a cabo de la siguiente manera:

Se introduce un patrón de los del conjunto de aprendizaje, perteneciente, por ejemplo, a la clase A, se obtiene la salida que genera la red para dicho ejemplo. Si el output producido es 1, la respuesta de la estructura es correcta y no se realizará ningún cambio en el valor de los parámetros. Si por el contrario, la salida producida es -1 se comete un error de clasificación y la red categoriza a este

patrón en B, produciéndose el aprendizaje y por ende la modificación en el valor de los pesos.

2.4.2. Adaline

Este modelo se diferencia del Perceptron en que sus salidas no son binarias sino reales, por lo que pueden modificar cualquier tipo de salida y se convierten en sistemas de resolución generales. Pueden resolver problemas a partir de ejemplos en los que es necesario aproximar una función cualquiera *F(x)*, definida por un conjunto de inputs. Los datos de entrenamiento en este caso son conjuntos de valores enteros compuestos por un valor de entrada y su salida asociada:

$$P = \{(x_1, y_1), ..., (x_m, y_m)\} \tag{2.7}$$

En este caso la función a aproximar sería:

$$F(\vec{x_i}) = y_i \qquad \forall p \in P \tag{2.8}$$

En otros términos, habrá que buscar la función $F(x)$ tal que aplicada a cada una de las entradas x_i del conjunto de aprendizaje P produzca la salida y_i correspondiente.

El aprendizaje en el Adaline incluye la diferencia entre el valor real producido en la capa de salida para un patrón de entrada y el que debería haber producido, es decir, su salida esperada contenida en el conjunto de aprendizaje. Con lo anterior no es posible conseguir una salida exacta, pero sí minimizar el error cometido por la red para la totalidad del conjunto de patrones de ejemplo. La manera de reducir esta medida de error global es recurrir a un proceso iterativo en el que se presentan los patrones uno a uno, y se van modificando los valores de los pesos mediante la *regla del descenso del gradiente o regla Delta*, en cuyo desarrollo se profundizará posteriormente.

2.4.3. Diferencias entre Adaline y Perceptron

- En el Perceptron la salida es binaria, en el Adaline es real.
- En el Perceptron la diferencia entre entrada y salida es 0 si ambas pertenecen a la misma categoría y ± 1 si por el contrario pertenece a categorías diferentes. En el Adaline se calcula la diferencia real entre entrada y salida.
- En el Adaline existe una medida de cuanto se ha equivocado la red, en el Perceptron sólo se determina si se ha equivocado o no.
- En el Adaline hay una razón de aprendizaje para regular cuanto va a afectar cada equivocación a la modificación de los pesos. Es siempre un valor entre 0 y 1 para ponderar el aprendizaje.

2.5. Fase de aprendizaje en una RNA

La parte más importante de una RNA es el aprendizaje pues es lo que determina el tipo de problemas que será capaz de resolver. Las RNA son sistemas de aprendizajes basados en ejemplos o patrones (*inputs*). La capacidad de una red para resolver un problema estará ligada de forma fundamental al tipo de ejemplos de los que dispone en el proceso de aprendizaje. Desde el punto de vista de los ejemplos, el conjunto de aprendizaje debe poseer las siguientes características:

- Ser significativo. Debe haber un número suficiente de ejemplos.
- Ser representativo. Los componentes del conjunto de aprendizaje deberán ser diversos. Si se tienen mucho más ejemplos de un tipo que del resto, la red se especializará en dicho subconjunto de datos y no será de aplicación general. Es importante que todas las regiones significativas del espacio de estados estén suficientemente representadas.

El aprendizaje consiste en la determinación de los valores precisos de los pesos para todas las conexiones, que la capacite para la resolución eficiente de un

problema. El proceso general consiste en ir introduciendo paulatinamente todos los ejemplos del conjunto de aprendizaje, y modificar los pesos de las conexiones siguiendo un determinado proceso. Una vez introducidos todos los ejemplos se comprueba si se ha cumplido cierto criterio de convergencia; de no ser así se repite el algoritmo y todos los ejemplos del conjunto vuelven a ser introducidos. La modificación de los pesos puede hacerse después del ingreso de cada ejemplo del conjunto, o una vez introducidos todos ellos.

El criterio de convergencia depende del tipo de red utilizado o del problema a resolver. La finalización del periodo de aprendizaje se puede determinar:

- Mediante un número fijo de ciclos. Se decide a priori cuántas veces será introducido todo el conjunto, y una vez superado dicho número se detiene el proceso y se da por aceptada la red resultante.

- Cuando el error descienda por debajo de una cantidad preestablecida. Se define en primer lugar una función para el error ya sea de manera individual para cada patrón o bien para la totalidad del conjunto de entrenamiento. Posteriormente se establece un valor aceptable para la variación y el proceso solo se detiene cuando la red produzca un error por debajo del valor prefijado. Para este criterio puede suceder que jamás se consiga estar por debajo de dicho nivel, en cuyo caso se debe disponer de un criterio adicional de parada, por ejemplo un número determinado de ciclos. En este caso la red se dice que no ha sido capaz de obtener una solución y será necesario probar cambiando alguno de los parámetros.

- Cuando la modificación de los pesos sea irrelevante. Se utiliza cuando se aplican modelos en los que los pesos de las conexiones se modifican con menor intensidad. Si el proceso de aprendizaje continúa, llegará un momento en que ya no se producirán variaciones en los valores de los

pesos de ninguna conexión; en este momento se dice que la red ha convergido y se detiene el proceso de aprendizaje.

Dependiendo del esquema de aprendizaje y del problema a resolver, se pueden distinguir tres tipos de esquema de aprendizaje:

1. Aprendizaje supervisado. Los datos del conjunto poseen dos atributos: los datos propiamente dichos y cierta información relativa a la solución del problema. Cada vez que un ejemplo es introducido y se procesa para obtener una salida, ésta se compara con el valor que debería haber producido, y de la que se dispone al estar incluida dicha información en el conjunto de aprendizaje. La diferencia entre ambas influirá en como se modificarán los pesos. Si los dos datos son muy diferentes, los pesos habrán de modificarse mucho y viceversa. Para este tipo de aprendizaje se dice que hay un profesor externo encargado de determinar si la red se está comportando en forma adecuada, mediante la comparación entre la salida producida y la esperada, y de actuar en consecuencia modificando apropiadamente los valores de los pesos.

2. Aprendizaje no supervisado. Los datos del conjunto de aprendizaje sólo tienen información de los ejemplos, por lo que la red modificará los datos a partir de información interna. Cuando se utiliza aprendizaje no supervisado, la red trata de determinar características de los datos del conjunto de entrenamiento: rasgos significativos, regularidades o redundancias. A este tipo de modelos también se les conoce como sistemas autoorganizados.

3. Aprendizaje por refuerzo. Es una variante del aprendizaje supervisado en el que no se dispone de información concreta del error cometido por la red para cada ejemplo, sino que simplemente se determina si la salida para dicho patrón es o no adecuada. Para este tipo de aprendizaje se tiene una

serie de características específicas que es importante resaltar. En este caso el conjunto de aprendizaje está compuesto por ejemplos que contienen los datos y sus salidas deseadas. El proceso consiste en modificar los pesos hasta que para todos los ejemplos del conjunto de entrenamiento, la salida producida sea lo más parecida posible a la deseada. Sin embargo, esto no siempre indica que la red sea capaz de solucionar el problema para valores futuros desconocidos.

Para poder determinar si la red produce salidas adecuadas o no se divide el conjunto de entrenamiento en dos conjuntos, el primero llamado de *entrenamiento* y el segundo de *validación*. El conjunto de entrenamiento se utiliza para aprender los valores de los pesos. La diferencia es que en vez de medirse el error en el conjunto de entrenamiento, se utiliza el de validación. De esta manera, para medir la eficacia de la red para resolver el problema, se utilizarán datos que no han sido utilizados para su aprendizaje. Si el error sobre el conjunto de validación es pequeño, entonces quedará garantizada la capacidad de generalización de la red.

Para que este proceso sea eficaz, ambos conjuntos deben tener las siguientes características:

- El conjunto de validación debe ser independiente del de aprendizaje.

- El conjunto de validación debe cumplir las propiedades de un conjunto de entrenamiento.

Además el conjunto de validación puede utilizarse durante el aprendizaje para guiarlo en conjunción con el de entrenamiento, en este caso el proceso sería el siguiente:

1. Asignar a los pesos valores aleatorios.

2. Introducir todos los ejemplos del conjunto de entrenamiento, modificando los pesos de acuerdo con el esquema de aprendizaje supervisado elegido.

3. Introducir todos los ejemplos del conjunto de validación. Obtener el error producido al ingresar dichos ejemplos.

4. Si el error calculado en el paso anterior está por encima de cierto valor umbral, ir a (2).

5. Acabar el proceso de aprendizaje y dar como salida la red obtenida.

2.6. Perceptron multicapa

La arquitectura del Perceptron multicapa se caracteriza porque tiene sus neuronas o nodos agrupados en capas o estratos de diferentes niveles. Cada una de ellas está formada por un conjunto de células y se distinguen tres tipos de capas diferentes: la de entrada, las ocultas y las de salida. Las neuronas de entrada se encargan únicamente de recibir las señales o patrones que proceden del exterior y propagar dichas señales a los nodos del siguiente estrato. La última capa actúa como salida de la red, proporcionando la respuesta para cada uno de los patrones de entrada. Las neuronas de los estratos intermedios realizan un procesamiento que es en la mayoría de los casos no lineal de los ejemplos recibidos.

Las conexiones del Perceptron multicapa siempre están dirigidas hacia adelante y llevan asociado un número real, llamados pesos, de la misma manera todas la neuronas de la red también tiene asociados un umbral, que en el caso del modelo analizado suele tratarse como otra conexión más cuya entrada es constante e igual a 1. Generalmente todos los nodos de una capa están conectados a todos los de la siguiente. (*Ver figura 2.2*)

2.6.1. Expresiones para calcular las activaciones de las neuronas de la red

Sea un perceptron multicapa con *C* capas, *C-2* capas ocultas y n_c neuronas en la capa *c*, para *c*= 1,2,…, *C*. Sea $W^c = (w_{ij}^c)$ la matriz de pesos asociada a las conexiones que van de la capa *c* a la *c+1* para *c*= 1,2,…, *C-1*, donde w_{ij}^c representa el peso de la conexión de la neurona *i* de la capa *c* a la neurona *j* de la capa *c+1*; y sea $U^c = (u_i^c)$ el vector de umbrales de las neuronas de la capa *c* para *c*= 2, 3,…, *C* . Se denota (a_i^c) a la activación de la neurona *i* de la capa *c* las cuales se calculan del siguiente modo:

- Activación de las neuronas de la primera capa.

$$a_i^c = x_i \text{ para } i = 1,2,\ldots,n_1 \quad (2.9)$$

- Activación de las neuronas de las capas ocultas.

$$a_i^c = f\left(\sum_{j=1}^{n_c} w_{ji}^{c-1} a_j^{c-1} + u_i^c\right) \text{ para } i = 1,2,...,n_c \text{ y } c = 2,3,...,C-1 \quad (2.10)$$

- Activación de las neuronas de la capa de salida.

$$y_i = a_i^C = f\left(\sum_{j=1}^{n_{C-1}} w_{ji}^{C-1} a_j^{C-1} + u_i^C\right) \text{ para } i = 1,2,...,n_C \quad (2.11)$$

Donde $Y = (y_1, y_2, ..., y_{nC})$ es el vector de salida de la red.

La función f es la llamada función de activación. Para el Perceptron multicapa las funciones de activación más utilizadas son la sigmoidal y la tangente hiperbólica. Dichas funciones tienen como imagen un rango continuo de valores dentro de los intervalos $[0,1]$ y $[-1,1]$, respectivamente. Ambas son funciones con dos niveles de saturación: el máximo, que proporciona salida 1, y el mínimo, salida 0 en la primera y -1 para la segunda. Las dos funciones están relacionadas mediante a expresión $f_2(x) = 2 f_1(x) - 1$, por lo que la utilización de una u otra se elige únicamente en función del recorrido que interese.

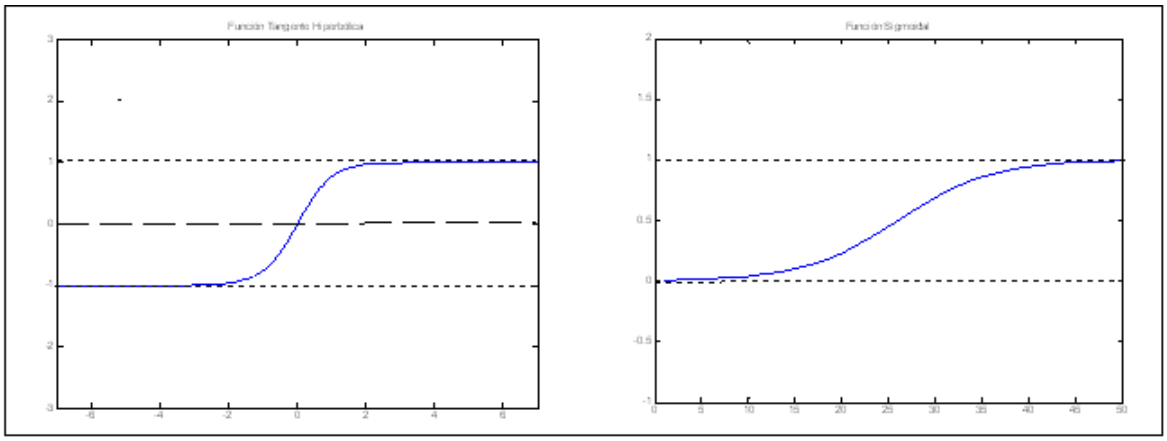


Figura 2.3. *Funciones de Activación más comunes utilizadas por las redes neuronales artificiales.*

Por las ecuaciones dadas anteriormente se observa que el Perceptron multicapa se define a través de sus conexiones, nodos y una función continua no lineal del espacio $\mathbf{R}^{n1}$ (espacio de los patrones de entrada) al espacio $\mathbf{R}^{nC}$ (espacio de los patrones de salida). Se puede escribir por tanto que:

$$Y = F(X, W) \tag{2.12}$$

Donde W es el conjunto de todos los parámetros de la red y F es una función continua descrita por las ecuaciones mencionadas. Dada esta característica se le reconoce como un "aproximador universal", pues tendrá la capacidad de aproximar cualquier tipo de función, incluso no continuas.

2.6.2. Capacidad de aproximación de funciones continuas

Una red neuronal de 2 capas ocultas con un número suficiente de nodos, con tipo de función de activación sigmoidal y función lineal en la capa de salida es capaz de aproximar cualquier función continua $f : R^n \rightarrow R$ con la aproximación deseada.

El teorema de Kolmogorov concerniente a la realización arbitraria de funciones multivariadas provee un soporte teórico a las redes neuronales que implementan tales funciones.

Teorema de Kolmogorov. Cualquier función continua valuada en los reales $f(x_1, x_2, ..., x_n)$ definida sobre $[0,1]^n$, $n \geq 2$ puede ser representada de la forma:

$$f(x_1, x_2, ..., x_n) = \sum_{j=1}^{2n+1} g_j \left[\sum_{j=1}^{n} \Phi_{ij}(x_i) \right] \tag{2.13}$$

dónde los términos g_j son correctamente elegidos funciones continuas de una variable y Φ_{ij} son funciones continuas monótonamente crecientes independientes de f.

La idea básica del teorema es representar la red con la siguiente arquitectura, donde *M* aplica Φ a cada x_i y posteriormente son sumadas:

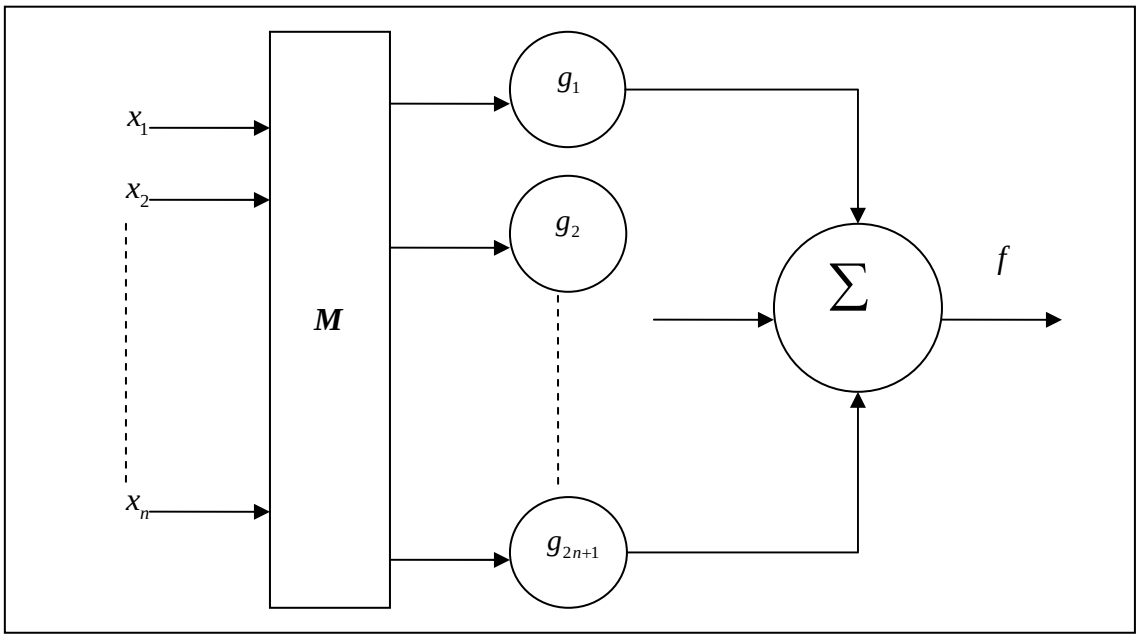


Figura 2.4. *Interpretación gráfica del Teorema de Kolmogorov.*

Teorema (Cybenko). Sea φ cualquier función sigmoidal continua, es decir, $\varphi(\xi) = \frac{1}{(1+e^{-\xi})}$. Dada cualquier función f en los reales sobre $[0,1]^n$, o algún otro subconjunto compacto de R^n, y $\varepsilon > 0$, existen vectores $w_1, w_2, ..., w_n, \alpha, \theta$ y una función parametrizada $G(\bullet, w, \alpha, \theta) : [0,1]^n \to R$ tal que

$$\left|G(x,w,\alpha,\theta)-f\right|<\varepsilon \quad \forall x\in[0,1]^n \tag{2.14}$$

donde

$$G(x,w,\alpha,\theta)=\sum_{j=1}^{N}\alpha_j\varphi(w_j{}^T x+\theta_j) \tag{2.15}$$

y $w_j\in R^n$, $\alpha_j,\theta_j\in R$, $w=(w_1,w_2,...,w_N)$, $\alpha=(\alpha_1,\alpha_2,...,\alpha_N)$ y $\theta=(\theta_1,\theta_2,...,\theta_N)$

De manera coloquial lo que el teorema de Cyenko señala es que cualquier función f puede ser ajustada por una función G constituida por una combinación lineal de funciones sigmoidales, que son precisamente las utilizadas por las redes neuronales.

2.6.3. Diseño de la arquitectura de un Perceptron multicapa

Cuando se aborda un problema utilizando un Perceptron multicapa, uno de los primeros pasos a realizar es el diseño de la arquitectura de la red, el cual implica la determinación de la función de activación a emplear, el número de neuronas y de capas. Existen diversas alternativas para identificar la estructura de una red neuronal, en este trabajo se abordaran tres de los métodos más utilizados dados sus buenos resultados: el algoritmo de retropropagación y el algoritmo genético, así como la mezcla de ambos. De la misma forma se ofrece un criterio para determinar el número de retrasos (sliding time window) que pueden optimizar el aprendizaje de la red, cuando se desea aplicar este modelo para el pronóstico de una serie de tiempo el cual se establece de manera heurística dados resultados empíricos.

2.6.3.1. Algoritmo de retropropagación

Puesto que el objetivo es que el output final de la red sea lo más próximo posible a la salida deseada, el aprendizaje se formula como un problema de minimización del siguiente modo:

$$Min_W E \tag{2.16}$$

Donde E es una función error que evalúa la diferencia entre las salidas de la red y las deseadas. En la mayor parte de los casos, ésta se define como:

$$E = \frac{1}{N}\sum_{n=1}^{N} e(n) \tag{2.17}$$

Donde N es el número de patrones y $e(n)$ es el error cometido por la red para el patrón n, definido por:

$$e(n) = \frac{1}{2}\sum_{i=1}^{n_c} (s_i(n) - y_i(n))^2 \tag{2.18}$$

De este modo, si *W** es un mínimo de la función error, en dicho punto el error es próximo a cero, lo cual implica que la salida de la red es próxima a la deseada, alcanzando así la meta de la regla de aprendizaje. Por lo tanto el aprendizaje del Perceptron es equivalente a encontrar un mínimo en la función error. La presencia de las funciones de activación no lineales hace que la respuesta de la red sea no lineal respecto a los parámetros ajustables, por lo que el problema de minimización es del mismo tipo y las técnicas de optimización que se usan también son no lineales y se basan en una adaptación de los parámetros siguiendo la dirección negativa del gradiente de la función *E* (Método de descenso del gradiente). No obstante, se han desarrollado varios métodos para localizar el mínimo de dicha función, y métodos basados en técnicas evolutivas, en los que la búsqueda está guiada por una función de adaptación.

Aunque el aprendizaje de la red debe realizarse para minimizar el error total, el procedimiento más utilizado, está basado en métodos del gradiente estocástico, los cuales consisten en una sucesiva minimización de los errores para cada patrón, $e(n)$, en lugar de minimizar el error total *E*. Por lo tanto, aplicando el método del gradiente estocástico, cada parámetro w de la red se modifica para cada patrón de entrada n de acuerdo con la siguiente ley de aprendizaje:

$$w(n) = w(n-1) - \alpha \frac{\partial e(n)}{\partial w} \tag{2.19}$$

Donde α es la razón o tasa de aprendizaje, parámetro que influye en la magnitud del desplazamiento en la superficie del error.

Debido a que las neuronas de la red están agrupadas en capas de distintos niveles, es posible aplicar el método de gradiente de forma eficiente, resultando el algoritmo de *retropropagación* o *regla delta generalizada.* El término retropropagación se utiliza debido a la forma de implementar el método del gradiente en el Perceptron multicapa, pues el error cometido en la salida de la red es propagado hacia atrás, transformándolo en un error para cada una de las neuronas ocultas.

2.6.3.2. Regla Delta Generalizada

Para el desarrollo de la regla Delta Generalizada es necesario distinguir dos casos:1) para los pesos de la capa oculta *C-1* a la de salida y para los umbrales de la capa de salida y 2) para el resto de los pesos y umbrales de la red.
Para 1):
Utilizando el método del gradiente el parámetro w_{ji}^{C-1} siguiendo la dirección negativa del gradiente del error:

$$w_{ji}^{C-1}(n) = w_{ji}^{C-1}(n-1) - \alpha \frac{\delta e(n)}{\delta w_{ji}^{C-1}} \qquad (2.20)$$

Por lo tanto, para la actualización de dicho parámetro es necesario evaluar la derivada del error $e(n)$ en dicho punto. De acuerdo con la expresión del error *(2.18)* y teniendo en cuenta, por un lado, que las salidas deseadas $s_i(n)$ para la red son constantes que no dependen del peso y, por otro lado, que el peso w_{ji}^{C-1} sólo afecta a la neurona de salida i, $y_i(n)$ se obtiene que:

$$\frac{\delta e(n)}{\delta w_{ji}^{C-1}} = (s_i(n) - y_i(n)) \frac{\delta y_i(n)}{\delta w_{ji}^{C-1}} \qquad (2.21)$$

A este punto se le tiene que calcular la derivada de la neurona de salida $y_i(n)$ respecto al peso w_{ji}^{C-1}. La salida de la red es la función de activación

f aplicada a la suma de todas las entradas por sus pesos respectivos. Aplicando la regla de la cadena para derivar la composición de dos funciones y teniendo en cuenta que, de todos los términos de la suma *(2.18)*, el único en el que interviene el peso w_{ji}^{C-1} es $w_{ji}^{C-1} a_j^{C-1}$ y por tanto, el único cuya derivada es distinta de cero, se obtiene:

$$\frac{\delta y_i(n)}{\delta w_{ji}^{C-1}} = f'\left(\sum_{j=1}^{n_{C-1}} w_{ji}^{C-1} a_j^{C-1} + u_i^C\right) a_j^{C-1}(n) \qquad (2.22)$$

Se define el término δ asociado a la neurona i de la capa de salida C y al patrón n, $\delta_i^C(n)$, del siguiente modo:

$$\delta_i^C(n) = -(s_i(n) - y_i(n)) f'\left(\sum_{j=1}^{n_{C-1}} w_{ji}^{C-1} a_j^{C-1} + u_i^C\right) \qquad (2.23)$$

Realizando las sustituciones correspondientes se obtiene:

$$\frac{\delta e(n)}{\delta w_{ji}^{C-1}} = \delta_i^C(n) a_j^{C-1}(n) \qquad (2.24)$$

Finalmente, reemplazando la derivada del error $e(n)$ respecto al peso w_{ji}^{C-1}, se obtiene:

$$w_{ji}^{C-1}(n) = w_{ji}^{C-1}(n-1) + \alpha \delta_i^C a_j^{C-1}(n) \qquad (2.25)$$

Para $j = 1, 2, ..., n_{C-1}$ $i = 1, 2, ..., n_C$

En el Perceptron multicapa el umbral de una neurona se trata como una conexión más a la neurona cuya entrada es constante e igual a 1. Se sigue de la ecuación anterior que los umbrales de la capa de salida se modifican de acuerdo con la siguiente expresión:

$$u_i^C(n) = u_i^C(n-1) + \alpha \delta_i^C \text{ para } i = 1, 2, ..., n_C \qquad (2.26)$$

Para 2):

Siguiendo el método del descenso del gradiente, la ley para actualizar dicho peso viene dada por:

$$w_{kj}^{C-2}(n) = w_{kj}^{C-2}(n-1) - \alpha \frac{\delta e(n)}{\delta w_{kj}^{C-2}} \tag{2.27}$$

En este caso el peso w_{kj}^{C-2} influye en todas las salidas de la red por lo que la derivada del error $e(n)$ respecto de dicho peso viene dada por la suma de las derivadas para cada una de las salidas, es decir:

$$\frac{\delta e(n)}{\delta w_{kj}^{C-2}} = -\sum_{i=1}^{n_C} (s_i(n) - y_i(n)) \frac{\delta y_i(n)}{\delta w_{kj}^{C-2}} \tag{2.28}$$

Para calcular la derivada de la salida $y_i(n)$ respecto al peso w_{kj}^{C-2} es necesario tener en cuenta que éste influye en la activación de la neurona j de la capa oculta $C-1$, a_j^{C-1}, y que el resto de las activaciones de las neuronas en esta capa no dependen de dicho peso. Por tanto se tiene que:

$$\frac{\delta y_i(n)}{\delta w_{kj}^{C-2}} = f'(\sum_{j=1}^{n_{C-1}} w_{ji}^{C-1} a_j^{C-1} + u_i^C) w_{ji}^{C-1} \frac{\delta a_j^{C-1}(n)}{\delta w_{kj}^{C-2}} \tag{2.29}$$

Haciendo las sustituciones respectivas se obtiene que:

$$\frac{\delta e(n)}{\delta w_{kj}^{C-2}} = \sum_{i=1}^{n_C} \delta_i^C(n) w_{ji}^{C-1} \frac{\delta a_j^{C-1}}{\delta w_{kj}^{C-2}} \tag{2.30}$$

Obteniendo la derivada de a_j^{C-1} con respecto a w_{kj}^{C-2}

$$\frac{\delta a_j^{C-1}}{\delta w_{kj}^{C-2}} = f'(\sum_{k=1}^{n_{C-2}} w_{kj}^{C-2} a_k^{C-2} + u_j^{C-1}) a_k^{C-2}(n) \tag{2.31}$$

Se define el valor δ para las neuronas de la capa $C-1$ como:

$$\delta_j^{C-1}(n) = f'(\sum_{k=1}^{n_{C-2}} w_{kj}^{C-2} a_k^{C-2} + u_j^{C-1}) \sum_{i=1}^{n_C} \delta_i^C(n) w_{ji}^{C-1} \tag{2.32}$$

Sustituyendo *(2.31)* y de acuerdo con el valor de $\delta_j^{C-1}(n)$ definido anteriormente, se obtiene que:

$$\frac{\delta e(n)}{\delta w_{kj}^{C-2}} = \delta_j^{C-1}(n) a_k^{C-2}(n) \tag{2.33}$$

Lo anterior implica la siguiente ley de aprendizaje:

$$w_{kj}^{C-2}(n) = w_{kj}^{C-2}(n-1) - \alpha \delta_j^{C-1}(n) a_k^{C-2}(n) \tag{2.34}$$

Para $k = 1, 2, ..., n_{C-2}$ y $j = 1, 2, ..., n_{C-1}$

Es posible generalizar la ley dado el resultado de la ecuación anterior para los pesos de la capa c a la capa $c+1$ ($c = 1, 2, ..., C-2$). Para ello basta tener en cuenta la activación de la que parte la conexión y el término δ de las neuronas de la siguiente capa. De este modo:

$$w_{kj}^{c}(n) = w_{kj}^{c}(n-1) - \alpha \delta_j^{c+1}(n) a_k^{c}(n) \tag{2.35}$$

Para $k = 1, 2, ..., n_c$, $j = 1, 2, ..., n_{c+1}$ y $c = 1, 2, ..., C-2$

Donde $\delta_j^{c+1}(n)$ viene dado por la siguiente ecuación:

$$\delta_j^{c+1}(n) = f'\left(\sum_{k=1}^{n_c} w_{kj}^{c} a_k^{c} + u_j^{c+1} \right) \sum_{i=1}^{n_{c+1}} \delta_i^{c+2}(n) w_{ji}^{c} \tag{2.36}$$

De igual manera se puede generalizar la ley de aprendizaje para el resto de los umbrales de la red, basta tratarlos como conexiones cuya entrada es constante e igual a 1.

$$u_j^{c+1}(n) = u_j^{c+1}(n-1) + \alpha \delta_j^{c+1}(n) \text{ para } j = 1, 2, ..., n_{c+1} \text{ y } c = 1, 2, ..., C-2 \tag{2.37}$$

Para la obtención de los valores de δ para *(2.19)* pertenecientes a cada neurona requiere el calculo de la derivada de la función de activación. Para la función sigmoidal y tangente hiperbólica se obtiene que:

- Función sigmoidal

$$f_1' = -\frac{1}{1+e^{-x}} \frac{e^{-x}}{1+e^{-x}} \tag{2.38}$$

Que cumple

$$f_1'(x) = f_1(x)(1 - f_1(x)) \tag{2.39}$$

Por lo anterior, si es utilizada esta función para obtener la activación de las neuronas de salida de la red los valores δ se obtienen de la siguiente manera:

$$\delta_i^C(n) = -(s_i(n) - y_i(n))y_i(n)(1 - y_i(n)) \quad (2.40)$$

Y los valores de δ para el resto de las neuronas de la red están dados por:

$$\delta_j^{c+1}(n) = a_j^c(n)(1 - a_j^c(n))\left(\sum_{i=1}^{n_{c+1}} \delta_i^{c+2}(n) w_{ji}^c\right) \quad (2.41)$$

Para $j = 1, 2, ..., n_{c+1}$ y $c = 1, 2, ..., C-2$

- Función tangente hiperbólica

Que cumple:

$$f_2'(x) = 2 f_1(x)(1 - f_1(x)) \quad (2.42)$$

Por lo que cuando se utilice esta función en vez de la sigmodal, los valores de δ para las neuronas de la red adoptan las expresiones anteriores multiplicadas por 2.

En el caso en que la función de activación de las neuronas de salida sea la función identidad, los valores δ para las neuronas de salida es:

$$\delta_i^C(n) = -(s_i(n) - y_i(n)) \text{ para } i = 1, 2, ..., n_C \quad (2.43)$$

A continuación se resumen las expresiones que definen la regla delta generalizada cuando se utiliza una función de activación sigmoidal, derivadas de la ecuación *(2.19)*:

- Pesos de la capa oculta *C-1* a la capa de salida y umbrales de la capa de salida.

Pesos:

$$w_{ji}^{C-1}(n) = w_{ji}^{C-1}(n-1) + \alpha\delta_i^C(n)a_j^{C-1}(n) \text{ para } j = 1, 2, ..., n_{C-1} \text{ y } i = 1, 2, ..., n_C$$

Umbrales:

$$u_i^C(n) = u_i^C(n-1) + \alpha\delta_i^C(n) \text{ para } i = 1, 2, ..., n_C$$

Donde

$$\delta_i^C(n) = -(s_i(n) - y_i(n))y_i(n)(1 - y_i(n))$$

- Pesos de la capa *c* a la *c+1* y umbrales de las neuronas de la capa *c+1* para $c = 1,2,...,C-2$.

Pesos:

$$w_{kj}^c(n) = w_{kj}^c(n-1) + \alpha\delta_j^{c+1}(n)a_k^c(n) \text{ para } k = 1,2,...,n_{C-1} \text{ , } j = 1,2,...,n_{c+1} \text{ y}$$

$$c = 1,2,...,C-2$$

Umbrales:

$$u_j^{c+1}(n) = u_j^{c+1}(n-1) + \alpha\delta_j^{c+1}(n) \text{ para } j = 1,2,...,n_{c+1} \text{ y } c = 1,2,...,C-2$$

Donde

$$\delta_j^{c+1}(n) = a_j^c(n)(1 - a_j^c(n))\sum_{i=1}^{n_{c+1}} \delta_i^{c+2}(n)w_{ji}^c$$

2.6.3.2. Razón de aprendizaje (α) y momento (η)

El parámetro α determina la magnitud del desplazamiento de los pesos, influyendo así en la velocidad de convergencia del algoritmo. Valores altos de la razón de aprendizaje, en principio podrían favorecer la convergencia más rápida, sin embargo, el método podría saltarse un mínimo o incluso puede oscilar alrededor de uno. Valores pequeños podrían evitar este problema, aunque reduzcan la velocidad de convergencia.

Para evitar la inestabilidad del algoritmo debido a la razón de aprendizaje se puede incluir un segundo término llamado *momento* denotada por η:

$$w(n) = w(n-1) - \alpha\frac{\partial e(n)}{\partial w} + \eta\Delta w(n-1) \qquad (2.44)$$

Donde $\Delta w(n-1) = w(n-1) - w(n-2)$ y η es un número positivo que controla la importancia asignada al incremento anterior. La inclusión de este nuevo término hace que la modificación actual del parámetro dependa de la dirección de la anterior, lo cual evita oscilaciones. Utilizando la ecuación anterior de manera recursiva se obtiene que:

$$\Delta w(n-1) = w(n-1) - w(n-2) = -\alpha \frac{\partial e(n-1)}{\partial w} + \eta \Delta w(n-2) = \\ ... = -\alpha \sum_{t=0}^{n-1} \eta^{n-1-t} \frac{\partial e(t)}{\partial w} \tag{2.45}$$

Por lo tanto la ley de la ecuación *(2.44)* puede escribirse de la siguiente forma:

$$w(n) = w(n-1) - \alpha \sum_{t=0}^{n} \eta^{n-t} \frac{\partial e(t)}{\partial w} \tag{2.46}$$

En esta expresión se observa que el cambio actual de un parámetro viene dado por la suma de los gradientes del error para todas las iteraciones anteriores. Por lo tanto, cuando la derivada parcial del error respecto al peso tiene signos opuestos en iteraciones consecutivas, la suma puede contrarrestar estos cambios de signo, y de este modo, procurar uno más suave en el peso, lo cual conduce a un método más estable. Por otra parte, si la derivada parcial del error respecto al peso tiene el mismo signo en iteraciones consecutivas, la utilización del momento procura un cambio mayor en el peso, acelerando así la convergencia del algoritmo.

2.6.4. Proceso de aprendizaje del Perceptron multicapa

Sea $\{(X(n), S(n)); n = 1, ..., N\}$ el conjunto de patrones que representan el problema a resolver, donde $X(n) = (x_1(n), ..., x_{n1}(n))$ son los ejemplos de entrada de la red, $S(n) = (s_1(n), ..., s_{nC}(n))$, son las salidas deseadas para dicho patrón y *N* es el número disponible de ellos.

Los pasos que componen el proceso de aprendizaje son los siguientes:

1. Se inicializan los pesos y umbrales de la red. Generalmente de forma aleatoria y con valores cercanos a cero.

2. Se toma un patrón *n* del conjunto de entrenamiento y se propaga hacia la salida de la red el vector de entrada $X(n)$ utilizando las ecuaciones *(2.9), (2.10)* y *(2.11)* obteniéndose así la respuesta de la red para dicho vector de entrada $Y(n)$.

3. Se evalúa el error cuadrático cometido por la red para el patrón *n* utilizando la ecuación *(2.17)*.

4. Se aplica la regla delta generalizada para modificar los pesos y los umbrales de la red. Para ello se siguen los siguientes pasos:

   4.1. Se calculan los valores δ para todas las neuronas de la capa de salida utilizando la ecuación *(2.23).*

   4.2. Se calculan los valores de δ para el resto de las neuronas de la red utilizando *(2.36)* empezando desde la última capa oculta y retropropagando dichos valores hacia la capa de entrada.

   4.3. Se modifican los pesos y los umbrales de la red siguiendo las ecuaciones *(2.25)* y *(2.26)* para los pesos y umbrales de la capa de salida y *(2.35), (2.37)* para el resto de los parámetros de la red.

5. Se repiten los pasos 2, 3 y 4 para todos los patrones de entrenamiento completando así una iteración o ciclo de aprendizaje.

6. Se evalúa el error total *E* cometido por la red.

7. Se repiten los pasos 2,3,4,5 y 6 hasta alcanzar un mínimo del error de entrenamiento, para lo cual se realizan *m* ciclos de aprendizajes.

2.6.4.1. Deficiencias del algoritmo de aprendizaje

- Mínimos locales. Debido a la utilización del método del gradiente para encontrar un mínimo en la función *E,* se corre el riesgo de que el proceso finalice en un punto extremo local, pues la condición $\frac{\partial E}{\partial w} \approx 0$ no garantiza que el mínimo alcanzado sea global. Una posible vía para evitar lo anterior es aumentar el número de neuronas ocultas de la red. En ocasiones, se considera que el proceso cae en un mínimo local debido a que la red posee un escaso poder de representación interna, de manera que no es capaz de distinguir diferentes patrones, proporcionando las mismas salidas para estos.
- Parálisis (saturación). Se produce cuando la entrada total a una neurona de la red toma valores muy altos, tanto positivos como negativos. Debido a que las funciones de activación poseen dos asíntotas horizontales, si la entrada a una neurona alcanza un valor muy elevado, ésta se satura, y alcanza una activación máxima o mínima. Por lo general, el fenómeno de parálisis ocurre cuando los valores de los parámetros de la red son elevados por lo que para evitar este problema es conveniente partir de valores iniciales aleatorios cercanos a cero.

2.7. Enfoque evolutivo del modelo de selección

El algoritmo genético fue presentado por Holland (1975) y desde entonces se han desarrollado diversas variantes computacionales. El modelo se explica como sigue: existen un número potencial de soluciones (*individuos*) para un determinado problema, que conforman una *población*; cada individuo es codificado por una cadena (*cromosoma*) de símbolos (*genes*) tomados de un alfabeto bien definido. Cada individuo tiene asignado un valor numérico (*adaptación*) que puede dar la solución apropiada. En cada generación, una fracción de la población es reemplazada por un nuevo conjunto de individuos generados por la aplicación de operadores genéticos, como la *mutación* y el *cruzamiento*, con el fin de crear

nuevas soluciones (*reproducción*).Todo el proceso es entonces evolutivo, donde los individuos mejor adaptados tienen mayor oportunidad de sobrevivir.

Los diferentes modelos de optimización evolutiva difieren principalmente en dos cosas: la función de adaptación y la representación de la red neuronal. Para la primera, el enfoque más usual es el considerar una medida de error sobre un conjunto independiente de validación. Por otro lado, la representación de dichas redes se ha diseccionado a dos alternativas: codificación directa e indirecta. El método directo engloba todos los detalles topológicos (pesos y conexiones), siendo este el más usado por su eficiencia y su fácil implementación. El segundo considera únicamente los parámetros más importantes o hace uso de la construcción de reglas.

Cuando se diseña un perceptron multicapa para el pronóstico de series de tiempo es común usar procedimientos de ensayo y error, como probar varias combinaciones de nodos ocultos. El uso de diseños evolutivos da una alternativa para realizar una selección topológica para las redes neuronales.

2.7.1. Algoritmos genéticos (GA) en la optimización de RNA

En su forma simple, el GA estándar es un método estocástico de optimización para problemas de programación discreta dado por:

$$\begin{aligned} &Maximizar \;\; f(s) \\ &s.a. \;\; s \in \Omega = \{0,1\}^n \end{aligned} \tag{2.47}$$

En este caso, $f : \Omega \to R$ es llamada función de adaptación o de activación, y los vectores binarios en Ω son llamados cadenas. En cada iteración se mantiene una colección de muestras provenientes del espacio de búsqueda más que de un simple punto.

Para comenzar el algoritmo en *(t=0)*, se crea una población inicial de M cadenas binarias denotadas por $S(0)=\{s_1,s_2,...,s_M\}\subset\Omega$, cada una con n bits. Usualmente esta población es creada de manera aleatoria porque no se conoce a priori en que lugar del espacio Ω se encontraran las cadenas que optimicen el problema. Si tal información es conocida se podría entonces, usar una población inicial cercana a la región a la que pertenezcan estos óptimos. A partir de esta población inicial, subsecuentes poblaciones $S(1),S(2),...,S(t),...$ pueden ser computadas empleando los tres operadores genéticos: selección, cruzamiento y mutación.

Selección

El GA estándar utiliza un método de ruleta giratoria para el operador de selección, el cual es una versión estocástica del mecanismo natural "el individuo mejor adaptado es el que sobrevive". Las cadenas en $S(t)$ que son candidatas para sobrevivir en la siguiente generación $S(t+1)$ se diseñan mediante una ruleta en la cual cada cadena perteneciente a la población es representada por una fracción de tal rueda de manera proporcional a su valor de adaptación con respecto a la función del total de la población.

Ejemplo:

Supongase que *M=5*, y considerese la siguiente población inicial de cadenas $S(0)=\{(10110),(11000),(11110),(01001),(00110)\}$, para cada cadena s_i en la población, la función de adaptación puede entonces ser valuada $f(s_i)$. La parte apropiada de la ruleta correspondiente a la *i-ésima* cadena se obtiene al dividir el valor de adaptación de ésta entre la suma de los valores de adaptación de la población entera:

$$\frac{f(s_i)}{\sum_{j=1}^{M} f(s_j)} \qquad (2.48)$$

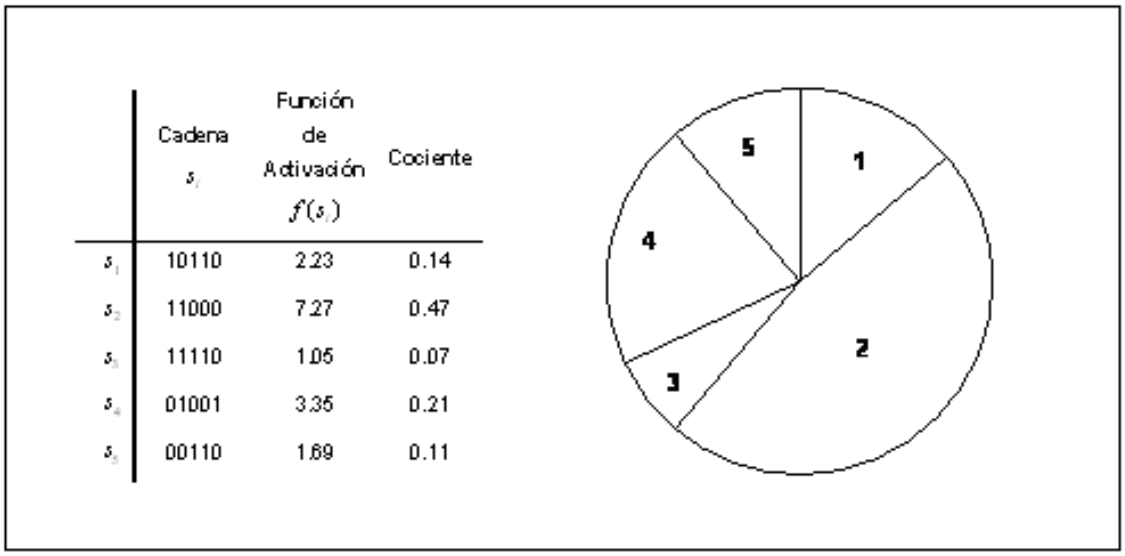


Figura 2.5. *Proceso de selección en el enfoque evolutivo.*

La figura 2.5 muestra el listado de las cadenas que componen la población y sus valores de adaptación correspondientes, así como la ruleta que ilustra el proceso de selección; los números en la ruleta corresponden al subíndice de la cadena a la que representan.

Después de asignar los valores de adaptación a cada cadena la ruleta es girada M veces, obteniéndose M cadenas. Es claro, que las cadenas (individuos) con mayor valor de adaptación tienen una mayor probabilidad de pertenecer a estos M individuos seleccionados.

El operador selección se aplica mediante el método de la ruleta giratoria, es decir, otorga cierta probabilidad a los individuos con menor valor de adaptación de pertenecer a la próxima generación, por la noción de que estos individuos pueden contener información parcial útil para la formación de la población subsecuente.

Cruzamiento

Antes de copiar los individuos obtenidos por selección a la siguiente generación se les debe de aplicar los otros dos operadores genéticos. Para el cruzamiento, que

es un mecanismo de recombinación, son elegidos pares de las M cadenas candidatas a pertenecer a la próxima población obtenidas previamente por el operador selección. La probabilidad de que tal operador sea aplicado será denotada por P_c, que de manera empírica se considera entre 0.6 y 0.99. El procedimiento es el siguiente:

1) Los pares de las cadenas seleccionadas son elegidas de manera aleatoria de $S(t)$ sin reemplazo.
2) Se elige un entero aleatorio k en $\{1, 2, ..., n-1\}$, llamado sitio de cruzamiento.
3) Los bits de las dos cadenas elegidas se intercambian después del k'ésimo lugar con una probabilidad de P_c.
4) Se repite el proceso hasta contemplar a toda la población $S(t)$.

De los tres operadores que participan en el proceso de búsqueda de los individuos óptimos, el cruzamiento es el operador crucial en la obtención de resultados globales, pues es el responsable de mezclar la información parcial contenida en ciertas cadenas de la población, razón por la cual P_c se considera elevado ya que se pretende abarcar la mayoría de los componentes del espacio de búsqueda.

Mutación

La mutación es un operador de complementación aplicado con probabilidad uniforme P_m, lo que significa que cada bit perteneciente a alguna cadena dentro de la población es intercambiado de 0 a 1 ó de 1 a 0 con una probabilidad de P_m.

El propósito de la mutación es diversificar la búsqueda e introducir nuevas cadenas en la población para explorar la totalidad del espacio de individuos. La creación de estos nuevas individuos es requerida dada la gran diferencia entre

el número de cadenas en la población M y el total del número de cadenas posibles en el espacio de búsqueda $\Omega, 2^n$. Típicamente M es elegida de magnitud pequeña comparada con 2^n, por lo que seleccionando y recombinando estas M cadenas en la población, sólo una fracción de Ω es explorada. Así pues, la mutación forza la diversidad en la población.

Sin embargo, el aplicar el operador mutación con gran frecuencia podría resultar en la destrucción de las cadenas con mayores valores de adaptación en la población lo que puede alentar e incluso impedir la convergencia hacia la solución. Tomando en cuenta la observación anterior es recomendable establecer que P_m se encuentre entre 0.01 y 0.001.[4]

Después de la mutación las cadenas candidatas a permanecer en tiempo $t+1$, son copiadas a la siguiente generación $S(t+1)$ y el proceso se repite evaluando los valores de adaptación para cada uno de los individuos y aplicando los tres operadores mencionados.

Para aplicar el algoritmo genético estándar a un problema arbitrario de la forma

$$\begin{aligned} &Maximizar \;\; y(x) \\ &s.a. \;\; x \in \Sigma \subset R^n \end{aligned} \tag{2.49}$$

es necesario establecer lo siguiente:

1. Una correspondencia entre el espacio de búsqueda Σ y algún espacio de cadenas binarias Ω, es decir, una relación de la forma $D: \Sigma \to \Omega$.
2. Una función de adaptación adecuada $f(s)$ tal que el máximo de f minimice a y.

---

[4] Bäck (1993) presentó un análisis teórico en el cual mostró que $P_m = \frac{1}{n}$ es el mejor valor cuando la función de adaptación es unimodal, no así cuando es multimodal, en donde aconseja que sea dinámica, es decir, que se inicialice en un valor $P_m(0)$ y decrezca hasta $\frac{1}{n}$.

2.7.2. Aplicación del GA a las RNA

La forma más obvia de aplicar el Algoritmo Genético al método de pronóstico que abarca este trabajo es para la búsqueda de los valores de los parámetros (pesos) óptimos en una red con una estructura definida (es decir, cuyo número de capas ocultas y nodos en cada una de ellas sea conocido). El uso de métodos de aprendizaje basados en el GA se justifican para tareas de aprendizaje que requieren la utilización de RNA con nodos ocultos, pues el GA es capaz de realizar búsquedas globales y es difícil de caer en mínimos locales. Además es muy útil para entrenar redes que consisten de unidades con función de activación no diferenciables, ya que la función de adaptación puede ser no diferenciable.

En el aprendizaje supervisado una forma de identificar una función de adaptación es $-E$, donde $E = E(w)$ puede ser la suma de los cuadrados de los errores.

Ejemplo:

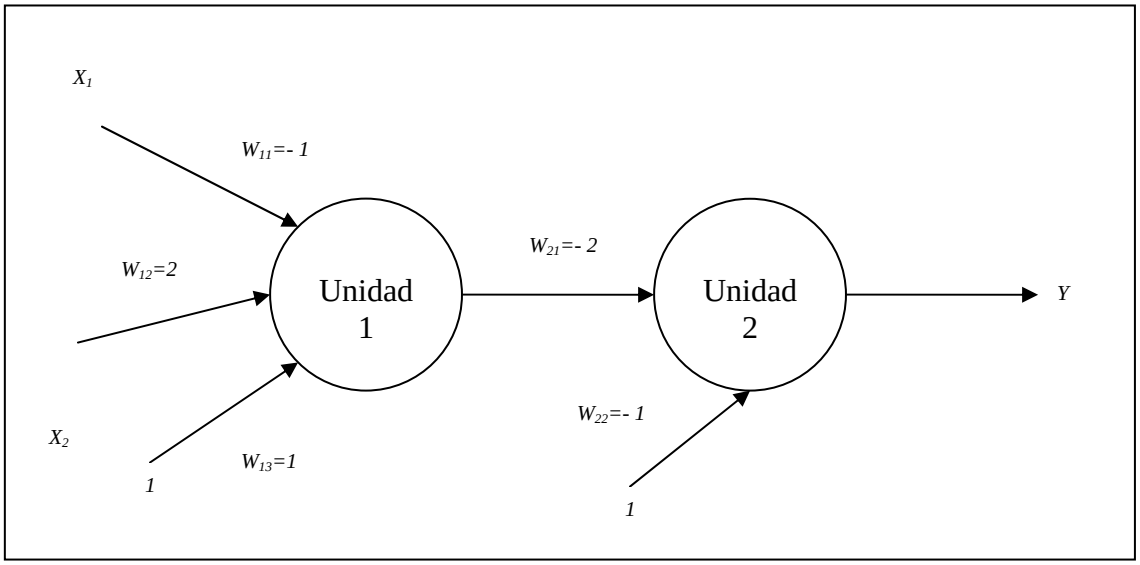


Figura 2.6. *Red compuesta de dos neuronas*.

En la figura 2.6 se muestra la estructura de una red con dos unidades neuronales. En este ejemplo cada peso se codifica con una subcadena de tres dígitos binarios que constituyen una cadena completa cuando son tomados en cuenta los pesos de todas las conexiones, es decir la estructura de una red. Aqui el primer bit de la

subcadena representa el signo del número binario que le sigue (110 representa -2 y 011 representa +3).

Ahora se puede generar una representación adecuada con el GA para la red como sigue **s**= (101,010,001,110,011) que corresponde a los valores reales de los pesos de la cadena $(w_{11}, w_{12}, w_{13}, w_{21}, w_{22})$ = (-1, 2, 1,-2, 3). Se comienza con una población aleatoria de tales cadenas (población de redes aleatorias), las generaciones sucesivas son construidas mediante el GA para evolucionar nuevos individuos. Las cadenas cuyo valor de adaptación sea superior al promedio de la población tienden a supervivir, y finalmente la población converge a la cadena mejor adaptada, es decir, a la red cuyos pesos producen un mínimo en la función de error.

2.7.3. Método Híbrido GA/ Método del descenso del gradiente para el entrenamiento de redes multicapas

Una de las ventajas de combinar el GA con el uso de redes neuronales artificiales se basa en la relación natural entre evolución y aprendizaje, pues la presencia de la primera facilita la segunda. En este contexto, el algoritmo genético es usado para mejorar el modelo evolutivo y el aprendizaje supervisado es a su vez utilizado para mejorar los mecanismos de aprendizaje.

Este método es de aprendizaje supervisado y es adecuado para redes neuronales interconectadas arbitrariamente. El objetivo es encontrar un conjunto apropiado de nodos ocultos que minimicen el error cometido por la red. A estas unidades de activación se les llama *hidden targets* porque estos pueden ser utilizados como vectores objetivo para entrenar a la primera capa.

El método del algoritmo genético en conjunción con el del descenso del gradiente (GA/GD) consiste en dos partes: el algoritmo genético se utiliza para la búsqueda

del conjunto de nodos apropiado y el uso del método de descenso del gradiente para encontrar el valor de los pesos en cada capa.

Considerándose una red totalmente interconectada con conjunto de entrenamiento $\{x^k, s^k\}, k = 1, 2, ...m$ (patrones y salidas deseadas respectivamente). Si se tiene un conjunto de nodos ocultos objetivo como vectores columna $\{h^1, h^2, ..., h^m\}$, $h^j \in \{0,1\}^J$ tal que son linealmente separables, entonces el método del descenso del gradiente puede ser utilizado para que de forma independiente y rápida se encuentren los pesos óptimos para las capas ocultas y la de salida. Inicialmente no se conoce el conjunto de nodos objetivo que resuelve de forma óptima el problema. Por lo tanto un GA es utilizado para explorar el espacio de dichos nodos objetivo y converger a una solución tal que $x \to h$ y $h \to s$. Dado que los hidden targets pueden ser valuados de forma binaria, es decir si existe el nodo entonces la variable buscada en el algoritmo genético toma el valor de 1 y cero en caso contrario, una forma natural de codificarlos es mediante una cadena de valores binarios $s = \{s^1, s^2, ..., s^m\}$ donde s^i es una cadena formada de los bits del vector h^i. Equivalentemente el espacio de búsqueda puede ser representado por un arreglo (matriz) $H = \left[h^1, h^2 ... h^m \right]$.

Una población de $\{H_j\}$ arreglos binarios $j = 1, ..., M$ es generada como la población inicial. Cada arreglo tiene asociado una etiqueta de su red correspondiente inicializadas con los mismos pesos. La función de adaptación de la *j*-ésima red H_j es determinada por su suma de cuadrados de los errores definida por :

$$E_j = \frac{1}{2} \sum_{k=1}^{M} \sum_{l=1}^{L} \left[d_l{}^k - (y_l{}^k)_j \right]^2 \qquad (2.50)$$

Donde $(y_l{}^k)_j$ es el *l*-ésimo output en la *j-ésima* red debido al input x^k.

Para inicializar el algoritmo GA/GD los pesos de los output de la capa de salida son adaptados sujetos al conjunto de entrenamiento $\{h^k, s^k\}$ de forma independiente de las capas precedentes. Después de que los pesos son obtenidos cada red (cadena) es entrenada y se computa su valor de adaptación. Posteriormente son utilizados los operadores genéticos para evolucionar a la siguiente generación de conjuntos de nodos objetivo $\{H_j\}$.

Durante la reproducción los $M/2$ conjuntos de nodos objetivo con los mejores valores de adaptación son duplicados y enteramente puestos en una población temporal a la cual se aplica el cruzamiento con una probabilidad P_c cercana a 1. Un par $\{H_i, H_j\}$ es seleccionado aleatoriamente sin reemplazo de la población temporal considerada. Si un patrón ha sido pobremente entrenado por la red i durante la fase de entrenamiento, es decir, el error es substancialmente grande al promedio del error cometido por el resto de las unidades de salida, entonces la columna correspondiente h^k de H_i es reemplazada por la *k'*esima columna de H_j. Aquí el cruzamiento puede afectar a múltiples pares de columnas en los arreglos de nodos objetivo.

La anterior es una descripción de un solo ciclo del algoritmo genético. Éste es repetido hasta que la población converge a una representación dominante o hasta que la medida de error es menor que un determinado valor.

Desventajas

El GA evoluciona los pesos basado en una medida de adaptación de la red completa (medida global de adaptación) y considera el estado resultante deseado. Ignorando la información del gradiente (o más generalmente la causa y efecto), cuando esta existe, el GA puede llegar a ser lento e ineficiente. Por lo anterior el algoritmo genera un gran costo en la velocidad y almacenaje al trabajar con poblaciones completas de redes, lo que lo vuelve impráctico para evolucionar diseños óptimos para redes de gran tamaño.

2.7.4. Modelo de Selección

Cuando se analizan las series de tiempo con redes neuronales, un punto crítico es la elección del modelo de selección, que se refiere al número de retrasos (*sliding time window)* a utilizar, así como a la estructura neurológica para la serie dada; ambas decisiones pueden transformar dramáticamente el resultado de los pronósticos. Un número de retrasos reducidos podría proveer escasa información a la red, mientras que uno grande puede aumentar la entropía, afectando el aprendizaje de la red, de la misma forma el uso excesivo o limitado de los nodos ocultos.

De manera heurística se pueden establecer cuatro reglas para generar una *slidig time window:*

1. Usar todos los datos atrasados <1,2,3,…,*m*> (*m* se considera por lo regular igual a 13 , valor considerado suficiente para modelar efectos estaciónales y de tendencia)
2. Usar todos los datos atrasados conteniendo valores autocorrelacionados sobre un umbral dado.
3. Usar los últimos cuatro datos con las mayores autocorrelaciones y;
4. Usar descomposiciones de la información dada, es decir,
    - <1, *K*, *K+1*> → si las series presentan efectos estaciónales (periodo *K*) y tendencia
    - <1, *K*> → si las series presentan efectos estacionales
    - <1> y <1,2> si la serie contiene tendencia

2.8. Uso de RNA para el pronóstico de series de tiempo

Predicción en un paso de tiempo

Consiste en predecir el valor de la serie en el instante de tiempo inmediatamente siguiente al instante actual t, a partir de las muestras disponibles hasta dicho instante de tiempo, es decir predecir el valor $x(t+1)$ utilizando un cierto número de muestras anteriores $x(t), x(t-1), x(t-2),...$ número que depende de la serie de tiempo.

Predicción en múltiples pasos de tiempo

Consiste en predecir el comportamiento de la serie en un futuro más lejano, concretamente en el llamado intervalo de predicción $[t+1, t+h+1]$, siendo h un número natural que representa el horizonte de predicción. Es decir, consiste en predecir los valores de la serie $x(t), x(t-1), x(t-2),...$ y $x(t+h+1)$ a partir de la información disponible en el instante de tiempo actual t.

2.8.1.Modelos neuronales para la predicción de series de tiempo

Para relacionar las observaciones de las series de tiempo con los patrones de entrada de las redes neuronales artificiales basta que la entrada de la red esté compuesta no sólo del patrón en un determinado instante de tiempo, sino también de una cierta historia de dicho patrón, la cual se representa utilizando una secuencia finita temporal en el pasado.

Mecanismos para la predicción de series de tiempo utilizando modelos de redes neuronales estáticas.

Predicción en un paso de tiempo. Considerando el vector $(x(t), x(t-1), x(t-2), ..., x(t-r))$ como patrón de entrada, las redes neuronales estáticas pueden utilizarse para aproximar F obteniendo el siguiente modelo de predicción:

$$\tilde{x}(t+1) = \tilde{F}(x(t), x(t-1), x(t-2), ..., x(t-r)) \quad (2.51)$$

Donde $\tilde{F}$ representa una aproximación neuronal de la función F y $\tilde{x}(t+1)$ es la predicción proporcionada por el modelo en el instante $t+1$.

Los patrones para el entrenamiento de la red se obtienen del conjunto de muestras disponibles de la serie temporal del siguiente modo: dado el conjunto $\{x(t)\}_{t=0,...,N}$ y el valor de r, cada patrón se construye utilizando los $r+1$ valores anteriores de la serie. De este modo cada entrada recoge la información temporal necesaria para predecir el valor de la serie en el instante de tiempo $t+1$, $\forall t = r, ..., N-1$.

Al realizar el entrenamiento, se pretende que la red capture la relación entre el valor de la serie en el instante de tiempo t y los r valores pasados. El ajuste de los pesos se realiza para minimizar el error cuadrático medio en la salida. Después del entrenamiento de la red, el modelo puede utilizarse con el propósito de predecir en un paso de tiempo; basta presentar a la red los valores de la serie en los $r+1$ instantes anteriores de tiempo.

Predicción en varios pasos de tiempo.

Esquema de predicción 1

Se utiliza de manera recurrente el modelo neuronal construido para la predicción en un paso de tiempo. Dicho modelo necesita como entrada los valores de la serie en los $r+1$ instantes de tiempo anteriores. Por lo tanto, si el objetivo es predecir el valor de la serie en el instante de tiempo $t+h+1$ la predicción viene dada por:

$$\tilde{x}(t+h+1) = \tilde{F}(x(t+h), x(t+h-1),..., x(t+h-r)) \quad (2.52)$$

Sin embargo, en el instante actual no toda la información de entrada a la red está disponible, pues los valores de la serie $x(t+h), x(t+h-1),..., x(t+1))$ para $h>1$**,** no se conocen. Para afrontar este problema se utilizan como entradas las salidas deseadas de la red en instantes anteriores de tiempo, en lugar de los valores de la serie temporal. El principal inconveniente de este esquema de predicción, es que los errores cometidos por los pronósticos son propagados hacia las predicciones futuras, pudiendo influir negativamente en la calidad del modelo.

Esquema de predicción 2

Consiste en construir un modelo NAR para predecir, directamente el valor de la serie en el instante $t+h+1$ a partir de la información disponible en el tiempo actual t :

$$x(t+h+1) = G(x(t), x(t-1), x(t-2),..., x(t-d)) \quad (2.53)$$

Donde G es una función no lineal que relaciona el valor de la serie en el instante $t+h+1$ con una sucesión finita de valores de la serie temporal disponibles en hasta el tiempo t**.** A diferencia del modelo para predecir en un paso de tiempo, en este caso el valor de la serie en el instante $t+h+1$ no se predice utilizando los $r+1$ valores inmediatamente anteriores, sino la información disponible en el momento actual.

Utilizando como entrada el vector $(x(t), x(t-1), x(t-2), ..., x(t-d))$, la red se puede utilizar para aproximar la relación G obteniendo el siguiente modelo de predicción:

$$\tilde{x}(t+h+1) = \tilde{G}(x(t), x(t-1), ..., x(t-d)) \quad (2.54)$$

Los patrones para el entrenamiento de la red se obtienen de las muestras disponibles de la serie temporal y se considera como salida en el instante de tiempo t, el valor de la serie en $t+h+1$. El aprendizaje de la red se realiza para minimizar el error medido en la salida en dicho instante; es decir:

$$e(t+h+1) = \frac{1}{2}(x(t+h+1) - \tilde{x}(t+h+1))^2 \forall t = d, ..., N-h-1 \quad (2.55)$$

La formulación de este modelo de predicción tiene sentido siempre que exista una relación, aunque desconocida a priori, entre el valor de la serie en el instante de tiempo $t+h+1$ y la secuencia $(x(t), x(t-1), x(t-2), ..., x(t-d))$. Hay que destacar también que a medida que aumenta el horizonte de predicción, las relaciones entre los valores temporales de la serie se pierden, y como consecuencia, no tendrá sentido planear esquemas de predicción de este tipo.

2.9. Aplicaciones de las RNA

A través de su evolución y gracias a herramientas cada vez más rápidas y eficientes que permiten el diseño de redes neuronales artificiales, éstas se han podido implementar exitosamente en diversas áreas teóricas y prácticas, entre ellas se tienen las siguientes:

Industria	Aplicaciones
Aeroespacial	Desempeño de habilidad aérea de pilotos automáticos, simulación de trayectoria de vuelo, habilidad aérea de sistemas de control.
Automotriz	Sistemas de dirección automáticos, garantía de análisis de actividad.
Bancaria	Evaluación de cheques y otros documentos.
Defensa	Discriminación de objetos, reconociendo facial, nuevos tipos de sensores, procesamiento de señales sonoras, radares y de imagen incluyendo comprensión de datos, extracción de características y supresión de ruido, identificación señal/imagen.
Electrónica	Predicción de secuencias de código, distribución de circuitos integrados de chips, control de procesos, análisis de deterioro de chips, síntesis de voz, modelación no lineal.
Entretenimiento	Animación, efectos especiales, pronóstico de mercado.
Finanzas	Estado de tasación real, recomendación de créditos, cobertura de hipotecas, calificación de bonos corporativos, análisis de uso de líneas de crédito, localización de actividad de tarjetas de crédito, programación de portafolios de inversión, análisis de finanzas corporativas, predicción de precios.
Industrial	Predicción de procesos industriales.
Seguros	Optimización de productos y aplicación de pólizas.
Manufacturera	Control de procesos de manufacturas, análisis y diseños de productos, diagnóstico de procesos de maquinarias, identificación de partículas en tiempo real, sistemas de identificación de calidad visual, análisis de diseño de productos químicos, análisis de mantenimiento de maquinaria, planeación y administración, modelación de la dinámica de sistemas de procesos químicos.
Médica	Análisis de células cancerígenas, diseño de prótesis, optimización de tiempo de transplantes, reducción de costos hospitalarios, mejoramiento de calidad hospitalaria.
Gasolina y gas	Exploración.
Robótica	Control de trayectoria, manipulador de controladores, sistemas de visión, robots de montacargas.

Lenguaje	Reconocimiento de lenguaje, comprensión de lenguaje, síntesis de texto para generar lenguaje.
Valores	Análisis de mercado, calificación de bonos automáticos, sistemas de identificación de stocks.
Industria	**Aplicaciones**
Telecomunicaciones	Comprensión de datos e imagen, servicios de información automática, translación de lenguaje hablado en tiempo real, sistemas de procesamiento de pago de clientes.
Transportación	Sistemas de diagnóstico de frenos de camiones, inventario de vehículos, sistemas de rutas.

Cuadro 2.1. *Aplicaciones de las Redes Neuronales Artificiales.*

Capítulo III
Inflación y política monetaria

Los bancos centrales son las autoridades responsables de proveer de moneda y de instrumentar la política monetaria. Esta última esta asociada al conjunto de acciones a través de las cuales la autoridad monetaria determina las condiciones bajo las cuales proporciona el dinero que circula en la economía.

La política monetaria no puede estimular en forma directa y sistemática a la actividad económica y al empleo por lo que la mejor contribución que puede hacer para fomentar el crecimiento económico sostenido es procurando la estabilidad de precios, pues un escenario de alta inflación desfavorece el ambiente económico, por tanto en años recientes los países con mercados emergentes han optado por seguir el ejemplo de los industrializados por adoptar regímenes de políticas monetarias con objetivos de inflación. México no es la excepción, y en 1995, se adhirió a este conjunto de países. Desde el 21 de enero de 2008, el objetivo operacional a través del cual el Banco de México realiza la instrumentación de la política monetaria se define para la tasa de interés interbancaria a un día ("tasa de fondeo bancario"). Al servir como guía para la instrumentación de la política monetaria, este instrumento le permite al Banco central comunicar la postura que tendrá sobre éste concepto al sector privado y al público en general.

La experiencia en países industrializados sugiere que dicho régimen debe contener las siguientes características para resultar exitoso en el control de inflación:

- Política fiscal fuerte y macroeconomía estable.
- Un buen desarrollo en el sistema financiero.
- Banco Central independiente y con poder para llevar a cabo la estabilidad de precios.

- Canales de transmisión[5] bien entendidos entre instrumentos políticos e inflación.
- Una fuerte y adecuada metodología para el pronóstico de inflación.
- Políticas transparentes que fomenten la credibilidad del Banco Central y el acceso a la información brindada por el mismo.

En este contexto se vuelve indispensable que el Banco Central cuente con un mecanismo adecuado y acertado para el pronóstico de inflación. La predicción de ésta variable es especialmente complicada para países emergentes. Los pronósticos de inflación juegan un papel primordial para conducir la política bajo objetivos de inflación pues existe un retrazo entre las acciones de índole monetario y su impacto al comportamiento de la inflación. Así pues podrían verse como un objetivo intermedio de la política monetaria. Los países emergentes poseen menos datos estadísticos para el pronóstico de este fenómeno debido a datos deficitarios, cambios estructurales en marcha o constantes y a su vulnerabilidad a desestabilizaciones económicas.

Se cuenta con diversos modelos a través de los cuales las instituciones centrales pueden establecer pronósticos para la variable en cuestión.

Los modelos macroeconómicos pueden facilitar la política monetaria en distintos modos. Aquellos a plazos cortos pueden ayudar a los bancos centrales a pensar en canales de transmisión. Más importante aún los modelos estadísticos pueden asistir en la presentación de pronósticos de inflación esenciales para la transparencia de información que debe promover el Banco Central.

Los modelos de series de tiempo poseen menos información económica acerca del fenómeno inflacionario que los macroeconómicos, pero proveen mejores pronósticos en el corto plazo. Los modelos univariados son relativamente fáciles

[5] Mecanismos a través de los cuales los bancos centrales pueden conocer los efectos que sus acciones tienen sobre la economía en general, y particularmente sobre el proceso de estabilización de precios

de llevar a cabo y permiten evaluar las condiciones económicas. Los multivariados permiten realizar el análisis de las interacciones entre otras variables y la inflación así como la inclusión de algunos supuestos estructurales.

3.1. Definición de Inflación

La inflación es el aumento sostenido y generalizado del nivel de precios de bienes y servicios, medido frente a un poder adquisitivo. Se define también como la caída en el valor de mercado o del poder adquisitivo de una moneda en una economía en particular[6].

En México el indicador al que hace referencia la inflación es el Índice Nacional de Precios al Consumidor (INPC) cuya finalidad es medir a través del tiempo la variación de los precios de una canasta de bienes y servicios representativa del consumo de los hogares del país.

3.2. Cálculo y componentes del INPC

Calculo del INPC

Un número índice es una medida de la proporción o porcentaje en el cambio de un conjunto de precios en un periodo de tiempo. Un índice de precios al consumidor (IPC) mide los cambios en los precios de bienes y servicios que se consumen en los hogares. Tales cambios afectan el poder adquisitivo de los ingresos de los consumidores y su bienestar. Como los precios de diferentes bienes y servicios no cambian con la misma tasa, un índice de precios sólo puede reflejar el promedio de este movimiento. A un índice típicamente le es asignado el valor de una unidad o 100 y son también usados para medir las diferencias en los niveles de precios entre distintas ciudades, regiones o países en un determinado periodo.

[6] Banco de México.

Índices Lowe

Una amplia y popular clase de índices de precios se obtiene definiéndolo como el porcentaje de cambio entre los periodos comparados en el total de los costos de adquisición de un conjunto de cantidades, generalmente llamado “canasta”. Esta clase de índice fue propuesto en 1823 por Lowe y se define de la siguiente manera:

Si se tienen n productos en la canasta con precios $p_j{}^t$ y cantidades $q_j{}^t$ en el tiempo t, el índice Lowe esta definido como sigue:

$$P_{Lo} = \frac{\sum_{i=1}^{n} p_i^t q_i}{\sum_{i=1}^{n} p_i^0 q_i} \qquad (3.1)$$

En principio cualquier conjunto de cantidades puede ser utilizado para la composición de la canasta, la cual ha de ser restringida a cantidades adquiridas en uno u otro de los dos periodos comparados o en cualquier otro lapso de tiempo. Por razones prácticas, la canasta de cantidades usadas para el cálculo del IPC se basa en el análisis del gasto de consumo de los hogares durante el periodo más cercano de los periodos cuyos precios son comparados.

El periodo cuyas cantidades son actualmente utilizadas es llamado *periodo de referencia de pesos* y se denota por la letra b. El periodo 0 es el periodo de referencia de precios, b puede ser cualquier periodo, incluyendo uno entre 0 y t, si el índice es calculado en algún momento posterior a t. Usando las cantidades del periodo b, el índice puede ser escrito de la forma:

$$P_{Lo} = \frac{\sum_{i=1}^{n} p_i^t q_i^b}{\sum_{i=1}^{n} p_i^0 q_i^b} = \sum_{i=1}^{n} (p_i^t / p_i^0)\, s_i^{0b} \qquad (3.2)$$

donde

$$s_i^{0b} = \frac{p_i^0 q_i^b}{\sum_{i=1}^{n} p_i^0 q_i^b}$$

El índice puede ser escrito y calculado de dos modos: con la proporción de los valores agregados, o como un promedio aritmético ponderado de las proporciones de los precios , o precios relativos, (p_i^t / p_i^0), para los productos individuales usando la parte de gasto hibrido s_i^{0b} como pesos llamados *híbridos* porque los precios y las cantidades pertenecen a los dos periodos, respectivamente 0 y b.

El cálculo del índice de precios en México utiliza una fórmula de ponderaciones fijas que lleva el nombre de “Laspeyres”, la cual consiste en calcular las medias ponderadas de los índices de los bienes y servicios que componen cada una de las agregaciones para las cuales se obtienen dichos cocientes, y se comparan con los calculados en el último periodo de medición. Por otro lado las ponderaciones permanecen fijas a lo largo del periodo de vigencia del sistema de índices de precios al consumo. El índice Laspeyres se deriva del Lowe cuando las cantidades pertenecen al periodo de referencia de precios, es decir, cuando $b = t$. Por lo tanto la expresión que lo representa es:

$$INPC = \frac{\sum_{i=1}^{n} p_i^t q_i^0}{\sum_{i=1}^{n} p_i^0 q_i^0} = \sum_{i=1}^{n} (p_i^t / p_i^0) s_i^0 \qquad (3.3)$$

donde s_i^0 denota la parte del gasto actual en el producto i en el periodo 0, esto es

$$\frac{p_i^0 q_i^0}{\sum_{i=1}^{n} p_i^0 q_i^0}$$

Por lo que no es sensible al hecho de que los consumidores introduzcan ajustes en su canasta cuando enfrentan cambios de precios relativos, nuevos productos, desaparición de otros, etc.[7] En general, el consumidor racional tiende a optimizar el costo de su canasta mediante este tipo de sustituciones.

Las principales críticas que se realizan al calcular un índice de precios de esta forma son las siguientes.

- Sesgo de sustitución: los índices de precios, al usar una canasta básica fija definida en el periodo base no toma en cuenta las substituciones de bienes que realizan los consumidores como respuesta a cambios de precios. Esta crítica se basa en la confusión de pretender que un IPC sea un Costo de Vida.
- No incorporan la introducción de nuevos bienes hasta que se efectúe una actualización en la cesta de productos.
- No incorpora cambios en la calidad.
- Pueden verse afectados los resultados si no se realizan debidamente las encuestas.
- No tiene en cuenta la economía informal, que en el caso mexicano juega un papel muy importante en el consumo de las familias.

Dado lo anterior, es importante llevar a cabo la actualización del INPC en dos sentidos, primero, realizar la modificación de año base de referencia de precios y segundo, llevar a cabo el cambio de base de ponderadores (referencia de cantidades).

El primer cambio se refiere a modificar el año a partir del cual se efectúan las comparaciones en las variaciones de los precios. El segundo, a modificar la matriz de los ponderadores respecto a los bienes y servicios representados en la canasta básica según su importancia, aparición, desaparición, etc.

[7] Considerar estos factores es de gran importancia si se toma en cuenta la apertura comercial de los países.

El Banco de México en el 2002 realizó la última actualización de los dos factores quedando establecido para el indicador actualizado lo siguiente:

- La nueva base de comparación es la segunda quincena de junio de 2002.
- La nueva base de ponderadores correspondiente es la estructura del consumo de los hogares observada en el 2000.[8]

Componentes

Como el índice que representa el nivel de precios en la economía mexicana, el INPC debe contener los bienes y servicios más representativos, refiriéndose a éstos como los más consumidos.

Con el propósito anterior el INPC se conforma de ocho grandes subgrupos constituidos de genéricos compuestos a su vez por productos específicos[9].

(1) Alimentos, bebidas y tabaco;
(2) Ropa, calzado y accesorios;
(3) Vivienda;
(4) Muebles, aparatos y accesorios domésticos;
(5) Salud y cuidado personal;
(6) Transporte;
(7) Educación y esparcimiento; y
(8) Otros servicios.

Para conocer el precio de tales productos y servicios el INEGI levanta la Encuesta Nacional de Ingresos y Gastos de los Hogares llevada a cabo de forma bianual[10].

[8] A partir de la segunda quincena de junio de 2002 los ponderadores y la canasta de bienes y servicios del INPC corresponden a la estructura del consumo de los hogares observada en el año 2000, pero actualizada a precios relativos a la quincena elegida como base de ponderación.
[9] Los productos específicos hacen referencia a marcas, presentaciones, modelos, etc. de cada uno de los bienes y servicios incluidos en el INPC. Los productos genéricos reagrupan a los productos específicos de forma más general en tipos de producto. Estos últimos constituyen la menor unidad de ponderación del INPC.

El gasto que reporta esta encuesta para cada bien o servicio se compara contra el gasto total que realizan las familias.

$$\omega_i = \frac{\text{Gasto en el bien o servicio } i \text{ de todas las familias mexicanas}}{\text{Gasto total de las familias mexicanas}}$$

el cociente resultante indica la importancia relativa o peso de cada uno de los bienes o servicios dentro del gasto total y al mismo tiempo establecen una correspondencia con los genéricos del INPC, por lo que puede determinarse el impacto que tendrá un cambio en el precio del genérico dentro del índice y por lo tanto dentro del presupuesto familiar.

3.3. Variables causantes de inflación

Para analizar el comportamiento de una variable económica, es importante no solo estudiar la variable en sí, sino también las relaciones que sostiene con otras, es decir, aquellos factores que intervienen en su comportamiento, y que según el tipo de análisis podrían ayudar a la comprensión de ésta.

Para los fines de este trabajo es conveniente mencionar qué variables pueden considerarse como causales de la inflación, pues posteriormente una de ellas será incluida para su pronóstico.

C.W. Granger (1969) propuso una definición de causalidad que es posible probar empíricamente.

Se puede decir que la variable Y causa a X si en la realización de un modelo para la observación del comportamiento de X la incorporación de Y reduce la

[10] Esta encuesta se levanta en 46 ciudades representativas ubicadas en siete distintas regiones del país cada dos años a partir del año 2002.

variabilidad de tal modelo, y la aumenta si no es incluida. De manera formal se puede resumir como sigue:

Sea A^t un proceso estocástico estacionario:

$$\overline{A^t} = \{A_{t-j},\ j = 1, 2, ..., \infty\}$$
$$\overline{\overline{A^t}} = \{A_{t-j},\ j = 0, 1, 2, ..., \infty\}$$
$$\overline{A(k)} = \{A_{t-j},\ j = k, k+1, ..., \infty\}$$

U_t *=toda la información en el universo acumulada hasta el momento* $t-1$

$U_t - Y_t$ *=toda la información contenida en* U_t *diferente de la serie de tiempo* Y_t

Definición. Causalidad. Si $\sigma^2(Y/U) < \sigma^2(Y/\overline{U-X})$ se dice que X causa a Y.

Definición. Retroalimentación. Si $\sigma^2(Y/\overline{U}) < \sigma^2(Y/\overline{U-X})$ y $\sigma^2(Y/\overline{U}) < \sigma^2(Y/\overline{U-X})$ se dice que X causa a Y.

Definición. Causalidad instantánea .Si $\sigma^2(Y/\overline{U},\overline{X}) < \sigma^2(Y/\overline{U})$ se dice que hay causalidad instantánea entre Y y X.

Definición. Causalidad retrazada. $X(t)$ causa a Y. *Se define la causalidad con retrazo* m, m *entero, como el último valor de* k *para el cual*

$$\sigma^2(Y/U - X(k)) < \sigma^2(Y/X(k+1)).$$

En este sentido y según diversos estudios realizados por diferentes autores y técnicas matemáticas aplicadas para probar causalidad[11] se proponen variables causales de inflación categorizadas en cuatro grupos: variables externas, variables internas, variables monetarias y variables de producción/demanda.

El primer grupo está representado por los precios y tasas de interés externos. Dentro del segundo tipo de variables se encuentran los salarios, los precios agrícolas y las tasas de interés domésticas. Contenidas en el tercer grupo están

[11] José Dávila, Alain Ize y José Morales."Fuentes del proceso inflacionario en México: Análisis de Causalidad"

las variables relacionadas con el financiamiento de gobierno, crédito interno y los agregados monetarios (específicamente el M1). En la última clasificación se consideran variables como la capacidad instalada.[12]

Tipo de Variable	Variable	Dirección de Causalidad	Efectos
Externas	Precios Externos (P^E)	$P^E \rightarrow P$	Apertura Progresiva hacia el exterior
	Tasas de interés Externas (r^E)	$r^E \rightarrow P$	Apertura Progresiva hacia el exterior
Costos Internos	Salarios (S)	$S \rightarrow P$	Una aceleración en los salarios es seguida de una desaceleración en los precios.
	Precios Agrícolas (P^A)	$P^A \leftrightarrow P$	Juegan un papel estabilizador en el proceso inflacionario
	Tasas (r)	$r \rightarrow P$	Relación muy débil
Monetarias	Financiamiento (F)	$F \rightarrow P$	Relación muy débil
	Crédito Interno (CI)	$CI \rightarrow P$	Relación muy débil
	Agregados Monetarios (B)	$B \rightarrow P$	Causalidad retrazada superior a un año
	M1	$M1 \rightarrow P$	Causalidad casi instantánea
Producción/Demanda	Capacidad Instalada	$C \rightarrow P$	Causalidad casi instantánea

Cuadro 3.1. *Variables Causales de la inflación en México.*

3.4. Agregados monetarios

Los agregados monetarios se definen como el valor de mercado (el determinado por la oferta y la demanda) de una suma de activos líquidos.

[12] Por la complejidad del fenómeno inflacionario, las variables que se relacionan con este fenómeno son numerosas y de diferente índole a las mencionadas, sin embargo, estas no son citadas porque se ha probado, no tienen tanta repercusión.

Existen cuatro tipos:

- El agregado monetario M1 está compuesto por los billetes y monedas en poder del público, las cuentas de cheques en poder de residentes del país y los depósitos en cuenta corriente. Los billetes y monedas en poder del público se obtienen al excluir la caja de los bancos del total de billetes y monedas en circulación.
- M2 incluye a M1, a la captación de residentes en la banca y en las sociedades de ahorro y préstamo, a los valores públicos y privados en poder de residentes y a los fondos para el retiro.
- M3 incluye a M2, a la captación bancaria de residentes del exterior y a los valores públicos en poder de residentes del exterior.
- El agregado monetario M4 incluye, además de M3 a la captación de las sucursales y agencias de bancos mexicanos en el exterior provenientes de residentes en el exterior y de nacionales.

La inflación en el mediano y largo plazo tiene un origen monetario, suposición que se basa en la teoría cuantitativa del dinero[13](Ver cuadro 3.1). No así en el corto plazo cuyo comportamiento se atribuye principalmente a la variabilidad de las cotizaciones de los bienes y servicios participantes en el índice de forma estacional. Por lo cual es razonable esperar que los agregados monetarios contengan información útil para la descripción del comportamiento de la inflación.

Por otro lado se han realizado pruebas, con distintas definiciones de dinero como las consideradas con anterioridad (financiamiento de gobierno, crédito interno y los agregados monetarios), que han determinado a los agregados monetarios, específicamente el M1, como el más influyente en el fenómeno inflacionario, conclusión no sorprendente pues el agregado M1 es el que capta mejor la trayectoria del dinero en la economía mexicana. En este sentido cobra importancia

[13] Dicha teoría señala que el aumento en la cantidad de dinero ofertada genera presiones inflacionarias. Es decir, $M*V=P*Y$ siendo M la cantidad de dinero, V la velocidad del dinero, P el nivel de precios y Y el nivel de producto.

el saber qué tipo de relación mantienen la inflación y el agregado monetario M1 y que tipo de modelo matemático podría emplearse para su análisis conjunto.

Diversos autores[14] han estudiado la relación entre los agregados monetarios y la inflación para varios países y México, encontrando una relación de carácter asimétrico y no lineal entre estas dos variables, originada cuando la economía responde menos a estímulos de política positivos que a negativos y viceversa, haciendo que el supuesto de una relación lineal en la modelación de estas variables sea injustificable. Por lo que en este trabajo se propone la utilización de un modelo no lineal: las Redes Neuronales Artificiales.

3.5. Importancia de la inflación

Una de las principales razones por las que se realiza una medición lo más precisa posible de la inflación es porque se trata de un fenómeno económico nocivo.

La inflación es perjudicial por las siguientes razones:

- Daña la estabilidad del poder adquisitivo de la moneda nacional;
- Afecta el crecimiento económico al hacer más riesgosos los proyectos de inversión;
- Distorsiona las decisiones de consumo y ahorro;
- Propicia una desigual distribución del ingreso; y
- Dificulta la intermediación financiera.

En otras palabras, una inflación elevada debilita el crecimiento de la actividad económica y del empleo. Deteriora los ingresos reales, al aminorar el poder adquisitivo de los salarios y otros ingresos, acentuando la inequidad y empobreciendo a los más desprotegidos. Dificulta la planeación de los individuos y las empresas y, por ende, promueve una asignación ineficiente de los recursos.

[14] Paula Garda, 2006; Emiliano Basco, 2006.

Así, un entorno inflacionario es un claro obstáculo para el crecimiento económico y el bienestar de la sociedad.

3.6. Previsiones para la inflación y balance de riesgos en México para el 2010

El escenario base inflacionario que ha previsto el Banco de México para el 2010 y 2011 engloba las siguientes consideraciones sobre la economía estadounidense y los mercados globales:

1.-La economía de Estados Unidos seguirá recuperándose en los siguientes trimestres, como consecuencia de un clima financiero más optimista y las políticas fiscales implementadas. Sin embargo, a largo plazo este crecimiento se verá afectado por la lenta recuperación del mercado laboral y la pérdida en la riqueza de las familias.

2.-Se espera que la recuperación de los mercados financieros internacionales apoye al crecimiento de la economía global, aunque estos aún sean susceptibles a cambios repentinos y no se encuentren completamente normalizados.

Tomando en consideración los anteriores puntos el pronóstico de inflación para el caso mexicano en los subsecuentes dos años se resume como sigue:

Se considera un alza en la inflación general anual durante el 2010. Lo anterior obedece al efecto que tendrán sobre el índice de precios las modificaciones tributarias aprobadas por el Congreso de la Unión[15], a la realineación de precios de los energéticos con sus referencias internacionales y los aumentos que determinan gobiernos locales. Cuyo efecto se espera se revierta en el 2011. Para

[15] Entre los cambios al sistema tributario se encuentran el incremento en las tasas del Impuesto Sobre la Renta (ISR), el Impuesto al Valor Agregado (IVA), el Impuesto a los Depósitos en Efectivo (IDE) y el Impuesto Especial sobre Producción y Servicios (IEPS).

este último año se pronostica una trayectoria descendente la cual se fundamenta en:

a.- Las modificaciones de carácter tributario tienen un carácter transitorio sobre la tasa de inflación.

b.-El restablecimiento de la política de deslices en los precios de los energéticos durante 2010 coadyuvará a la reducción del indicador.

c.-Los ajustes previstos para 2010 en los precios y derechos que determinan los gobiernos locales deben traducirse en un uso más eficiente de los recursos con que cuenta la economía. En la medida en que una parte importante de los ajustes se implemente durante este año el próximo podría observarse un menor ritmo de crecimiento en dichos precios.[16]

No obstante las previsiones de la reducción inflacionaria, ésta se encuentra sujeta a diversos riesgos, entre los que destacan:

i) Un deterioro en las expectativas de inflación de largo plazo a consecuencia del repunte de la variable en cuestión que se anticipa, lo que podría generar que las empresas trasladen sus mayores costos a los precios al consumidor de bienes y servicios no afectados directamente.

ii) Una magnitud distinta a la estimada en los ajustes previstos a las cotizaciones de los bienes y servicios administrados y concertados.

iii) Traspaso a precios de las modificaciones tributarias que aprobó el Congreso de la Unión distinto del estimado. Ello, debido a la incertidumbre asociada a todo tipo de estimación de parámetros mediante métodos econométricos.

[16] La medida de imponer topes a los precios de determinados productos a pesar de tener efectos reductores de inflación en el corto plazo no funcionan de la misma manera en un horizonte más amplio dado que reduce drásticamente los márgenes de ganancia de los sectores controlados, lo que implica que las empresas del estado tienen que estar subsidiadas. Por otra parte los sectores no subsidiados sufren de caídas en los márgenes de ganancia que tenderán a convertirse en caídas en la producción

iv) Condiciones climatológicas que afecten la oferta de frutas y verduras.

v) Una recuperación de la actividad económica a un ritmo distinto al que se anticipa sin duda incidiría en la trayectoria de la inflación. Si ésta fuese más lenta ello atenuaría las presiones inflacionarias, y lo contrario ocurriría si la reactivación fuese más rápida.

vi) Reversión de flujos de capital que incida adversamente sobre el tipo de cambio.

Capítulo IV

Pronóstico de inflación con RNA

La información utilizada en el documento corresponde a la inflación mensual reportada por el Banco de México de junio de 2002, cuando se llevó a cabo el cambio de bases en el índice a la primera quincena de marzo del presente año, último dato dado a conocer por la institución hasta la fecha en la que se desarrolla éste trabajo.

Con el fin de realizar el entrenamiento y validación de las redes se ha dividido el conjunto de patrones disponibles para cada determinado número de rezagos en dos, la primera mitad se utilizó para la fase de entrenamiento y la segunda para la validación del modelo como se muestra en la siguiente tabla:

Número de Retrasos	Horizonte de Pronóstico	Número total de patrones	Periodo de Entrenamiento	Periodo de Validación
4	6 meses	77	Jun 2002- Ago 2005	Sep 2005-Jun2009
	12 meses	71	Jun 2002- May 2005	Jun 2005-Jun2009
6	6 meses	75	Jun 2002- Jul 2005	Ago 2005-Jun2009
	12 meses	69	Jun 2002- Abr 2005	May 2005-Jun2009
12	6 meses	69	Jun 2002- Abr 2005	May 2005-Jun2009
	12 meses	63	Jun 2002- Ene 2005	Feb 2005-Jun2009

Tabla 4.1. *Fases de entrenamiento en la serie de tiempo utilizada.*

La evaluación de la capacidad de pronóstico en las distintas fases para los modelos estimados por RNA se realiza por medio de dos medidas de evaluación del error de pronóstico: la media de los cuadrados del error y la suma del cuadrado de los errores. Estos estadísticos miden la desviación de los valores pronosticados de inflación respecto a los observados en el periodo de análisis. La primera de ellas se aplica sobre cada una de las dos fases de aprendizaje y la segunda se toma sobre la totalidad de los conjuntos de patrones.

Con la finalidad de elegir la mejor *sliding time window* se proponen 6 modelos para el pronóstico de inflación; con 4, 6 y 12 rezagos de la serie de tiempo y con horizonte de pronóstico de 6 y 12 meses para cada uno de ellos (modelos base), como se muestra en la tabla 4.2.

Se parte de un modelo autorregresivo de inflación simple (modelo base) y se prueba si la incorporación del agregado monetario M1 mejora la precisión del pronóstico lo que da lugar a 6 modelos con el número de rezagos y horizontes de pronóstico iguales a los primeros pero con la inclusión de la variable M1 (modelos alternativos). Para estos últimos se consideran los 24 rezagos mensuales anteriores de la serie del medio circulante, es decir, se agrega a los modelos base la información de los últimos dos años de M1. Se adoptan estos rezagos pues al realizarse otros ensayos con diferente número de rezagos se llegó a resultados poco satisfactorios, lo que apoya la idea generalmente aceptada, de que el agregado monetario M1 ejerce influencia a mediano plazo sobre la inflación.

Los modelos de RNA que se ensayan fueron desarrollados en Matlab combinando sus herramientas para redes neuronales y resolución de algoritmos genéticos.

Para cada modelo se asumen dos capas ocultas. Al no conocerse a priori la topología de la red neuronal que brinda la mejor aproximación a la serie, el número de nodos en cada capa se determina mediante la utilización del algoritmo genético, para el cual se establece una población inicial de 10 individuos, con 20 cromosomas cada uno (10 cromosomas representan el número de nodos para la primer capa y otros 10 la segunda). Las probabilidades de mutación y cruzamiento son 0.01 y 0.80 respectivamente.

La elección final de la mejor red depende de las diferentes medidas de evaluación de la bondad de ajuste para cada fase de aprendizaje.

A continuación se muestran la estructura y las medidas de error para los modelos base

Número de Retrasos	Horizonte de Pronóstico	Estructura de la Red[17]	MSE Entrenamiento	MSE Validación	SSE
4	6 meses	6,8,1	0.0059	0.0853	4.6180
	12 meses	6,6,1	0.0295	0.0659	3.8537
6	6 meses	6,6,1	0.0068	0.0796	5.5952
	12 meses	5,9,1	0.0080	0.2020	9.2330
12	6 meses	6,9,1	3.66528e-6	0.0651	2.6190
	12 meses	5,7,1	1.11611e-5	0.0918	3.0808

Tabla 4.2. *Estructura y medidas de error para los modelos base.*

Como puede apreciarse los modelos con 12 retrasos son aquellos con mejor aproximación para las observaciones destinadas al entrenamiento de la red y aunque no sea así para la fase de validación (en cuyo caso son los que contemplan sólo 4 rezagos), vuelven a tener el mejor ajuste al considerar la totalidad de los patrones en la red (ver SSE). Lo anterior puede explicarse con la información expuesta en el apartado anterior pues la inflación tiene un comportamiento similar en determinadas temporadas para todos los años y lo más conveniente, según los resultados previos, es tomar la información que contemple todo este ciclo, es decir, un año (las últimas 12 observaciones mensuales).

En la tabla 4.3 se exponen los resultados para los modelos alternativos

Número de Retrasos	Horizonte de Pronóstico	Estructura de la Red	MSE Entrenamiento	MSE Validación	SSE
4	6 meses	9,5,1	6.47873e-6	0.0648	3.0248
	12 meses	6,6,1	0.0066	0.0602	2.7315
6	6 meses	6,6,1	29414e-5	0.0761	5.6782
	12 meses	7,7,1	2.54057e-5	0.0645	3.2648
12	6 meses	6,5,1	0.0004	0.0639	3.3668
	12 meses	5,4,1	0.0038	0.0496	1.8743

Tabla 4.3.*Modelos alternativos.*

[17] Cada cifra separada con el signo de “coma” corresponde al número de nodos en la primera y segunda capa ocultas respectivamente y a la capa de salida.

Al observar los datos precedentes puede llegarse a la misma conclusión que en el análisis de la tabla 4.2. Con la incorporación del agregado M1, los modelos con 12 rezagos siguen siendo los que mejor ajustan el comportamiento de la inflación, tanto en el lapso de validación como en el error considerando la totalidad de los patrones, e incluso poseen una ventaja más, en este caso desde un punto de vista computacional, pues también son aquellos que contemplan un menor número de nodos ocultos, lo que reduce el tiempo de computo.

La tabla 4.4 muestra de manera resumida la comparación entre los modelos básicos y los alternativos

Número de Retrasos	Horizonte de Pronóstico	Número total de nodos ocultos (primera y segunda capas)		Cambio de las medidas de error con la inclusión de M1 (%)		
		Sin M1	Con M1	MSE Entrenamiento	MSE Validación	SSE
4	6 meses	14	14	167.23	31.64	52.67
	12 meses	12	12	346.97	9.47	41.08
6	6 meses	12	12	-97.69	4.60	-1.46
	12 meses	14	14	89.00	213.18	182.80
12	6 meses	15	11	-99.08	1.88	-22.21
	12 meses	12	9	-99.71	85.08	64.37

Tabla 4.4. *Comparación entre modelos básicos y alternativos.*

Es notoria la mejoría en el ajuste cuando se incluye la información del agregado monetario M1. Se minimiza el error durante la fase de validación para todos los modelos propuestos. En tres de ellos se obtiene un menor ajuste en los patrones de entrenamiento, sin embargo, al compararlos con el error total cometido por la red, éstos empeoran en un porcentaje muy bajo. Se sacrifica un poco el ajuste en los datos de entrenamiento, pero éste mejora para los utilizados en la validación y por ende también se mejora la capacidad predictiva de la red. Por otro lado, se

conservan el número de nodos utilizados e incluso en los dos últimos modelos se reduce.

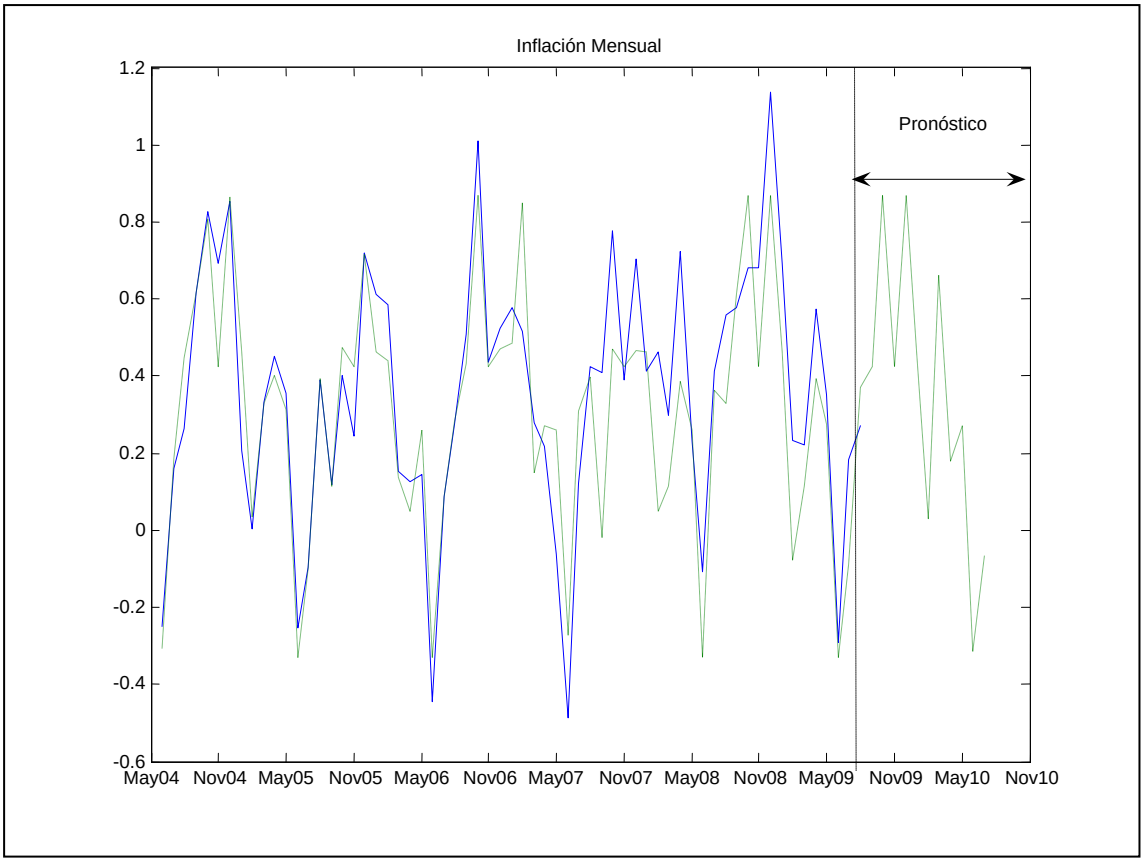
Con los resultados precedentes se puede concluir que el mejor modelo para pronosticar la inflación con redes neuronales artificiales es aquel que contempla los últimos 12 rezagos de la serie, 24 rezagos de la serie del medio circulante (M1) y posee dos capas ocultas, la primera con 5 nodos y la segunda con 4.

Dada la elección del mejor modelo pueden pronosticarse las próximas 12 observaciones de inflación

Mes	Pronóstico	Observada[18]	Desviación %
Ago-09	0.2247	0.2393	-6.0667
Sep-09	0.4784	0.5016	-4.6212
Oct-09	0.3398	0.3025	12.3232
Nov-09	0.5379	0.5187	3.6970
Dic-09	0.4776	0.4139	15.4040
Ene-10	0.9531	1.0870	-12.3232
Feb-10	0.6497	0.6585	13.8636
Mar-10	0.6334	0.7099	-10.7828
Abr-10	**0.3713**		
May-10	**-0.3135**		
Jun-10	**0.0675**		

[18] Datos posteriores a la realización de este trabajo que sirven para medir el ajuste de la red neuronal a datos no observados.

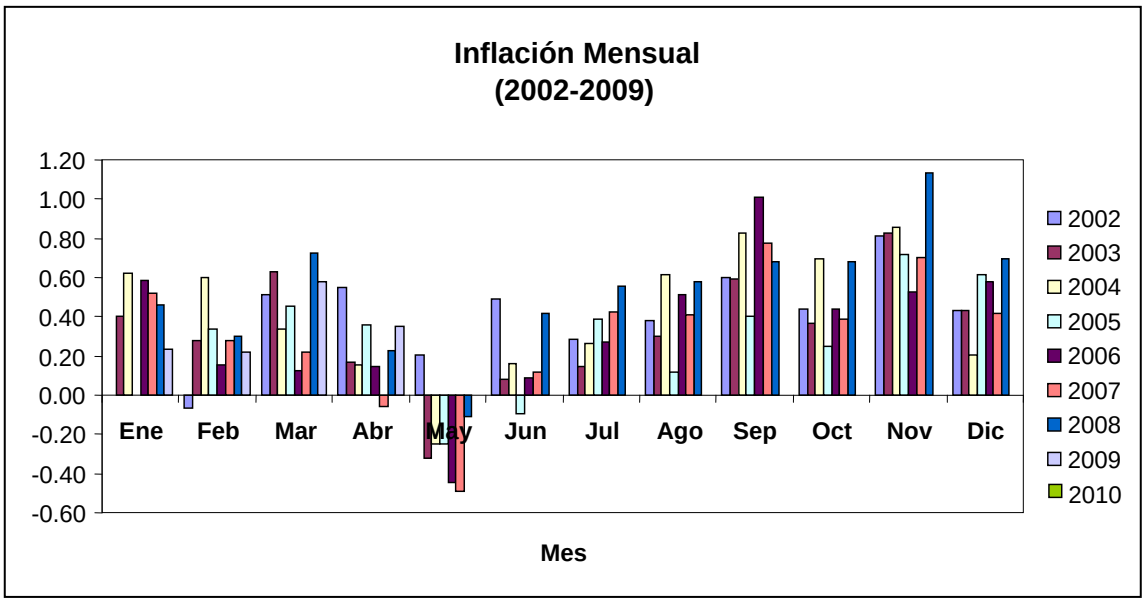
Los datos observados (línea continua), el ajuste (línea punteada) y datos pronosticados (después de la línea discontinua) se ilustran en la gráfica 4.1.



Gráfica 4.1. *Inflación observada y pronóstico.*

En la grafica anterior puede observarse que los pronósticos llevan una tendencia a la baja para los próximos meses, lo que concuerda con los objetivos de inflación del Banco de México para el año en curso y el siguiente, dadas las circunstancias económicas actuales.

La gráfica 4.2 ilustra el comportamiento de la inflación mes por mes desde el año de análisis, gracias a la cual se puede notar la sensibilidad que tiene el ajuste por RNA en cada uno de estos periodos.



Gráfica 4.2. *Comportamiento de la inflación mensual.*

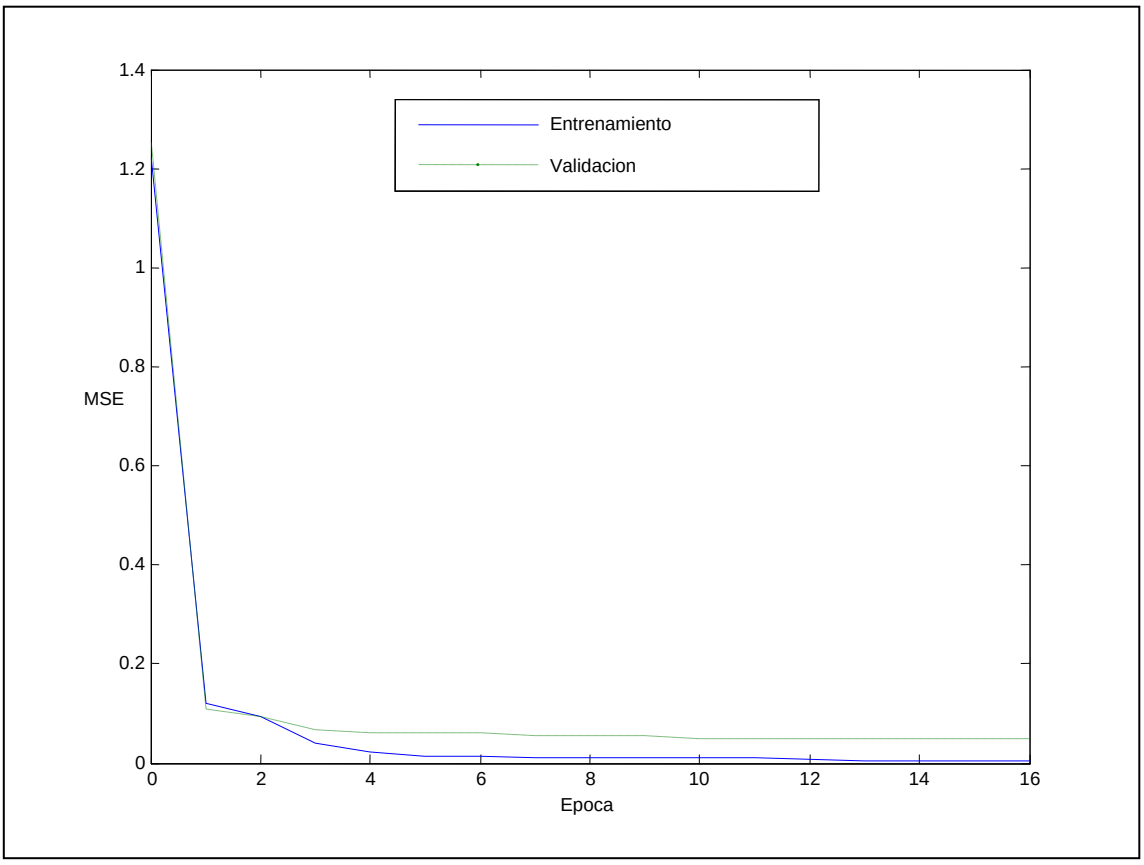
Principales variaciones

Durante el mes de mayo se presenta una disminución considerable en el indicador de inflación, la que se explica por la baja en las cotizaciones de los servicios y bienes administrados (gasolina, electricidad y uso de gas doméstico) y reducción en los precios de mercancías alimenticias (limón, cebolla, tomate verde, etc.).

El aumento en la inflación durante el mes de septiembre se debe a la mayor contribución al índice por parte del grupo de frutas y verduras, al aumento de la demanda de los bienes y servicios administrados, así como al alza en el precio de colegiaturas.

Las variaciones a la alza del mes de noviembre se explican principalmente por el repunte en los precios de productos agropecuarios y un importante aumento en el consumo de energía eléctrica.

A continuación se muestra la evolución de la medida de error MSE para las fases de entrenamiento en diferentes tiempos del aprendizaje de la red



Grafica 4.3. *Evolución de la medida de error MSE.*

Conclusiones

Considerando el efecto negativo de la inflación en las operaciones económico-financieras, es de vital importancia prever su comportamiento futuro para evitar dichas repercusiones.

El uso de modelos basados en RNA para el pronóstico de ésta variable (y para aquellas de carácter económico) presenta numerosas ventajas frente a otros métodos tanto en la reducción de tiempo en la modelación como ventajas técnicas pues no se requiere un tratamiento previo en la información. Al mismo tiempo permite la integración de otras variables en el análisis de la serie lo que produce un mejor ajuste a su comportamiento.

Al implementar el algoritmo genético en la búsqueda de la estructura óptima de la red a utilizar se minimiza el tiempo requerido en éste proceso y se garantiza la ejecución rápida del aprendizaje, sin embargo, tendrán que tenerse conocimientos previos de la variable estudiada para poder interpretar los movimientos sugeridos por el pronóstico obtenido, así como concientizarse de los elementos sociales y políticos que podrían influir en ella.

Actualmente existen diversos paquetes computacionales que incluyen aplicaciones de redes neuronales artificiales, por lo que, aunado al conocimiento extendido de su efectividad no solo en el pronóstico de series de tiempo sino en muchos otros y diversos campos, pueden llegar a facilitar su uso y promover su desarrollo.

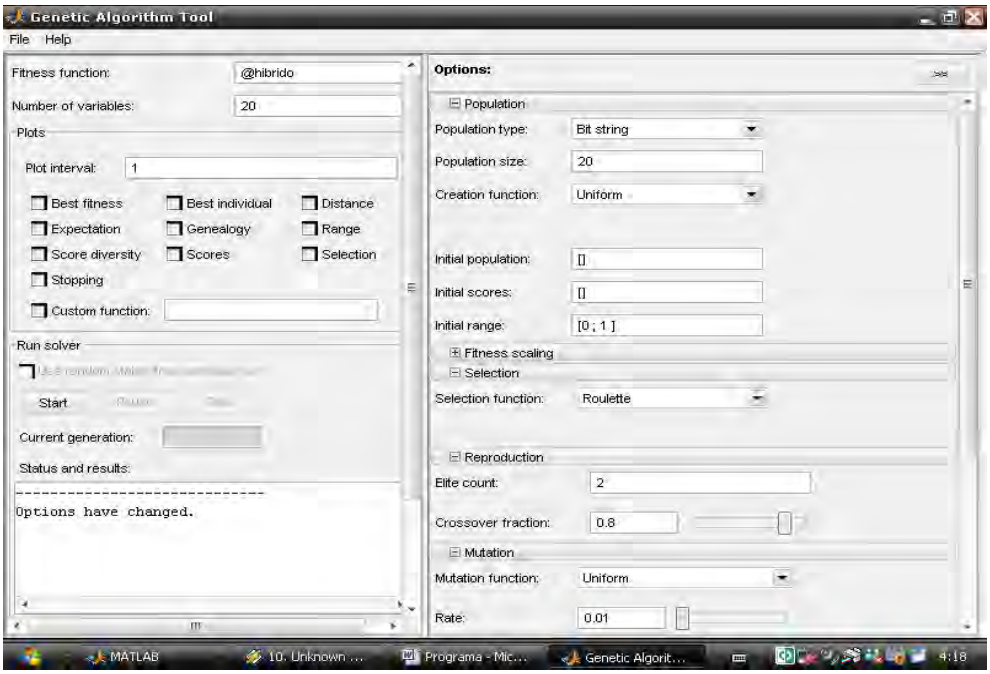
ANEXO I

Código en Matlab para la implementación del Algoritmo Genético en RNA, cuando se desconoce el número de nodos de la red para el pronóstico de la inflación en México.

Se toman datos mensuales de la inflación reportada por el Banco de Mèxico. Cada entrada de la red considera las últimas 12 observaciones (<X_{t-1}, X_{t-2}, X_{t-3}, X_{t-4}, X_{t-5}, X_{t-6}, X_{t-7}, X_{t-8}, X_{t-9}, X_{t-10}, X_{t-11}, X_{t-12}>) y las últimas 24 de la serie de variación mensual del agregado monetario M1(modelo final).La salida deseada corresponde a la observación de la serie en el tiempo *t* después de 12 meses, es decir X_{t+12}, por lo que la red es capaz de pronosticar datos de doce meses posteriores al ultimo dato reportado de inflación.

```
>> load inflacion_mensual.txt
>> %Definiendo los patrones%
>> inputs(inflacion_mensual,12)
>> load inputs
>> %Para la funciòn newff(que crea una red)en Matlab es necesario establecer%
>> %los rangos entre los que se encuentran los patrones%
>> rango_patrones(inputs)
>> load rango_patrones
>> %Definiendo las salidas deseadas%
>> target_n_steps(inflacion_mensual,12,12)
>> load targets
>> %cada salida deseada se pronostica en un intervalo de 12 meses%
>>%se hace uso de la herramienta gráfica de Matlab para la resolución del Algoritmo
Genético realizando los cambios convenientes, la función que despliega esta herramienta es
gatool %

>>gatool
```



Los cambios realizados son los siguientes:
Fitness Function: hibrido (Esta función será descrita posteriormente)
Number of Variables: 20 (es el numero de nodos sobre los cuales se aplica el algoritmo genético, 10 en cada una de las capas ocultas)
Population Type: Bit String (Las variables consideradas son de tipo binario)
Population Size: 10
Selection Function : Roulette
Las probabilidades de mutación y cruzamiento se definen 0.01 y 0.8 respectivamente.

A continuación se presentan las funciones utilizadas para encontrar las salidas deseadas, los patrones de entrada, y la matriz de rangos utilizadas para la elaboración de la red, de la misma forma se muestra la función *hibrido*, la cual busca minimizarse mediante el algoritmo genético.

```
function target_n_steps(x,d,s)
%cuando s=1=>caso, x=vector original de la serie de tiempo
n=length(x);
targets=zeros(1,n-d-s+1);
for i=1:n-d-s+1
   targets(i)=x(i+d+s-1);
end
save targets;
```

```
function inputs(x,R)
%x=vector original de la serie de tiempo
n=length(x)-R+1;
inputst=zeros(n,R);
 for j=1:n
     for i=0:R-1
        inputst(j,i+1)=x(i+j);
     end
 end
 inputs=inputst';
      save inputs
```

```
function rango_patrones(x)
%x es la matriz de inputs
n=length(x(:,1));
rango_patrones=zeros(n,2);
for i=1:n
    rango_patrones(i,:)=minmax((x(i,:)));
end
save rango_patrones
```

```
function z=hibrido(x)
%inflaciòn=vector de rangos de inputs

inflacionrangos=[];

%definir el vector de inputs

inflacion=[];
%definir el vector de salidas deseadas
inflaciontargets=[];

sizeLayer=10;%definir el numero de neuronas para los layers. debe ser el mismo para
todos
net=newff(inflacionrangos,[sum(x(1:sizeLayer)) sum(x(sizeLayer+1:2*sizeLayer)) 1]);
net = train(net,inflacion,inflaciontargets);
Y = sim(net,inflacion);
e=inflaciontargets-Y;
z=sse(e)
```

%Para la estructura de la red que resulta del algoritmo genético %

```
load inflacion_mensual.txt
clear inputs
clear targets
clear rango_patrones
```

```
clear net
rango_patrones(inputs)
load rango_patrones
target_n_steps(inflacion_mensual,12,12)%cambiar%
load targets
inflacionrangos=rango_patrones;

ntargets=length(targets);
inflacion=inputs(:,1:ntargets);

inflaciontargets=targets;

>> iitr = 1:round(length(targets)/2);
>> iival = round(length(targets)/2)+1:length(targets);

val.P = inflacion(:,iival);
val.T = inflaciontargets(:,iival);
ptr = inflacion(:,iitr);
ttr = inflaciontargets(:,iitr);
net=newff(inflacionrangos,[5 4 1]);%cambiar%
[net,tr]=train(net,ptr,ttr,[],[],val);
plot(tr.epoch,tr.perf,tr.epoch,tr.vperf)
legend('Entrenamiento','Validacion',-1);
ylabel('MSE'); xlabel('Epoca')
tr.perf
tr.vperf

Y = sim(net,inflacion);
e=inflaciontargets-Y;
z=sse(e)

-----------------------------------------

IF=inputs(:,ntargets+1:length(inputs));
YF = sim(net,IF)
Y2=[Y YF];
x=1:length(Y2);
%crea la matriz de los targets mas Nan en los valores para los pronosticos%
Y3=[inflaciontargets NaN(1,length(Y2)-length(inflaciontargets))];
plot(x,Y3,x,Y2)
```

% Resultados para el modelo que mejor ajusta el comportamiento de inflación % (Inflación Mensual , Retrasos:12, Pronostico 12 meses con agregados)

Population =

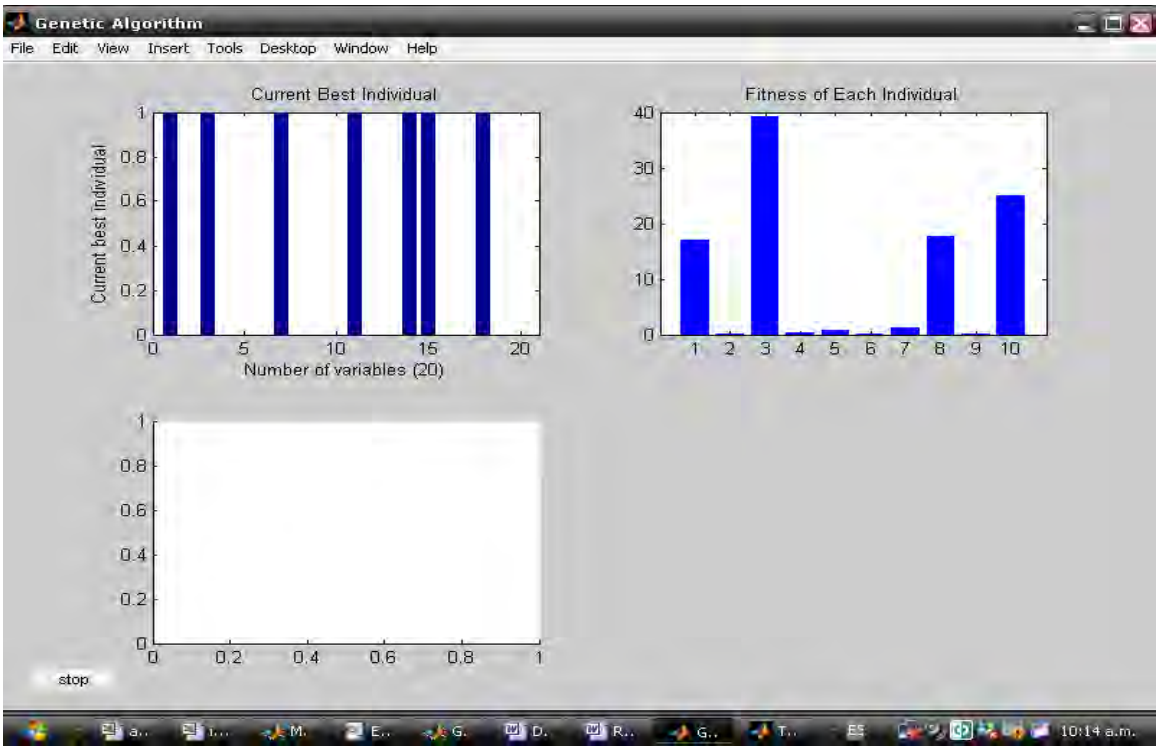
Columns 1 through 12

0	1	1	0	0	0	0	0	1	1	1	1
1	0	1	0	0	0	1	0	0	0	1	0
1	0	1	0	0	1	1	0	0	0	1	1
0	0	0	1	1	0	0	0	0	0	0	1
1	0	1	1	1	0	1	1	0	0	1	0
1	0	0	0	0	0	1	0	0	0	0	0
1	0	0	1	1	0	1	1	0	0	1	0
1	0	0	0	0	1	1	0	1	1	0	0
1	1	1	1	0	0	0	0	0	1	0	0
1	0	1	1	1	0	0	1	0	0	1	0

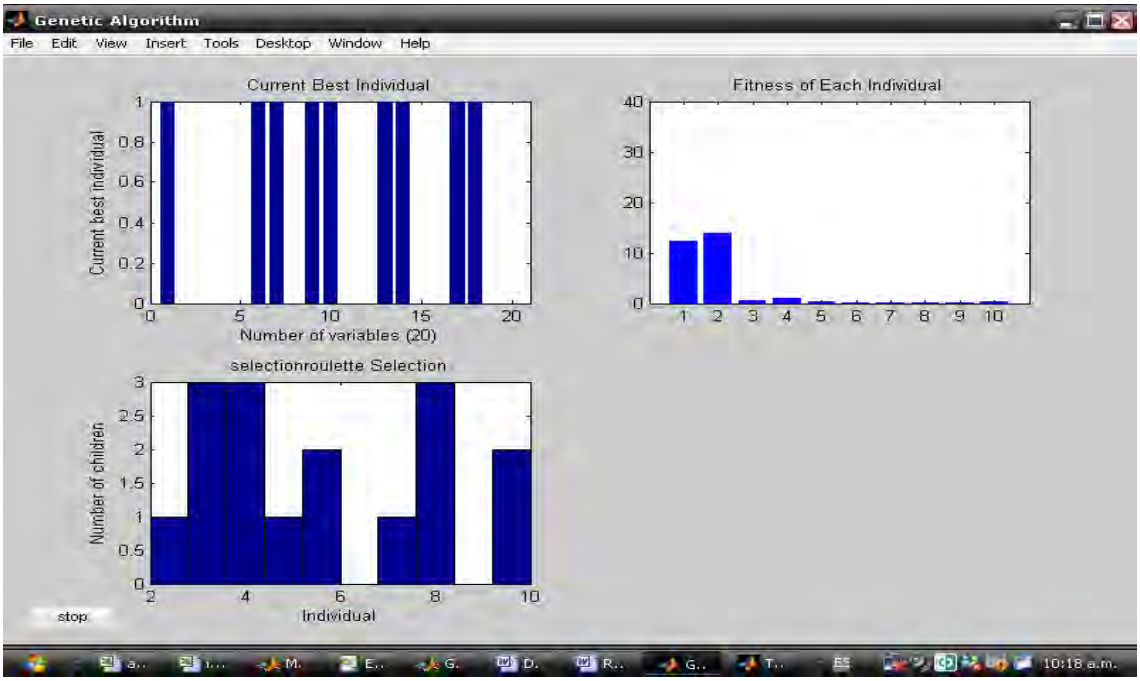
Columns 13 through 20

1	0	1	1	0	0	0	0
0	1	1	0	0	1	0	0
1	0	1	0	1	1	0	1
0	1	0	1	1	1	0	0
1	1	0	0	0	0	1	0
1	1	0	0	0	1	0	0
1	1	0	1	0	1	0	1
1	1	0	0	1	1	1	1
0	1	1	1	0	1	0	0
1	0	1	1	0	1	0	1

Primera Generación



Segunda generación



Fitness function value: 0.01888888645801403
Optimization terminated: stall time limit exceeded.

%Entrenamiento de la red optima%

TRAINLM, Epoch 0/100, MSE 1.2206/0, Gradient 79.3751/1e-010
TRAINLM, Epoch 16/100, MSE 0.00382444/0, Gradient 0.273286/1e-010
TRAINLM, Validation stop.

MSEentrenamiento=
Columns 1 through 7
1.2206 0.1195 0.0930 0.0400 0.0217 0.0139 0.0128
Columns 8 through 14
0.0113 0.0109 0.0107 0.0106 0.0097 0.0064 0.0055
Columns 15 through 17
0.0052 0.0044 0.0038

MSEvalidacion =
Columns 1 through 7
1.2517 0.1076 0.0942 0.0672 0.0601 0.0601 0.0601
Columns 8 through 14
0.0553 0.0552 0.0538 0.0496 0.0496 0.0496 0.0496
Columns 15 through 17
0.0496 0.0496 0.0496

SSE =

1.8743

%Pronosticos%

YF =

Columns 1 through 7
0.2247 0.4784 0.3398 0.5379 0.4776 0.9531 0.6586
Columns 8 through 11
0.6334 -0.3135 -0.0675 0.0675

Bibliografía

- **Lawrence,Fawset.** Fundamentals of Neural Networks. New York: Prentice Hall.
- **Rabañan, Juan.** Artificial Neural Networks in Real Life Applications. Idea Group Publishing.
- **Hassoun,Mohamad**. Fundamentals of Artificial Neural Networks. Cambridge, Mass. : MIT Press, c1995.
- **Alain,Ize Gabriel.** La Inflación en México.
- **Guerrero, Victor.** Análisis Estadístico de Series de Tiempo Económica. Fondo de Cultura Económica.
- **Banco de México.** Informe de inflación y política monetaria 2010-2011.
- **Banco de México.** Reportes Mensuales de Inflación.

- www.banxico.org.mx
- www.inegi.org.mx

- Ensayos varios sobre la inflación , Banco de México.