Obtención de Perceptrones multicapas de mínimo tamaño usando métodos de regularización.

J.A. Horas[(1)(2)(4)], J.P. Galanzini[*(1)(4)] y C. Kavka[(3)(4)].

Instituto de Matemática Aplicada (IMASL)[(1)]. Departamento de Física[(2)]. Departamento de Informática[(3)]. Facultad de Ciencias Físico, Matemáticas y Naturales[(4)]. Universidad Nacional de San Luis. Av. Ejercito de los Andes 950. San Luis. CP 5700.
email: jhoras@unsl.edu.ar, pgalanzi@unsl.edu.ar, ckavka@unsl.edu.ar.

Estudiamos la aplicación de una función objetivo regularizada en redes multicapas entrenadas por Backpropagation a fin de lograr una Red Neuronal de tamaño mínimo. Se compara la convergencia de las redes así obtenidas para varias funciones de regularización. Se muestran resultados para problemas de clasificación que tienen diverso grado de dificultad para ser aprendidos. Se estudia también la existencia de una zona óptima de aplicación del proceso de regularización en el espacio de las conexiones.

To obtain a neural network of minimum size we applied a regularized objective function in multilayer perceptron trained with backpropagation algorithm. The convergence is compared among diverse regularization functions. We show results for various classification problems. We also study the existence an optimal zone in which the regularization process must be applied.

I. Introducción

Los procedimientos de aprendizaje en Redes Neuronales realizan inferencia inductiva, esto es particularmente cierto cuando se entrenan Perceptrones Multicapas con el algoritmo de Backpropagation (BP). Se sabe que este puede interpolar aceptablemente bien, pero extrapola incorrectamente[(1)]. Moderar esto es posible a través de la determinación del número mínimo de neurodos en la capa oculta[(2,3)]. El número de neurodos en las capas de entrada y salida en cambio, esta fijado por el problema a resolver.

La realización de gradiente descendiente tiende además a ser "oportunista", depende del lugar de comienzo en el espacio de pesos y hace un uso innecesario y costoso de los parámetros disponibles. Es importante entonces el diseño de procedimientos para detectar y evitar esto último, cuyos efectos sobre la performance de generalización son devastadores.

Para aliviar este problema, en este trabajo analizaremos un método de regularización que adiciona a la función de costo un término que contabiliza la salida de cada neurodo oculto. Este es minimizado simultáneamente con el error cuadrático medio (MSE), lográndose así que algunos de dichos neurodos minimicen también su salida y puedan ser eliminados, sin perturbar la performance de la Red.

Este término de regularización permite, en principio, no solo suprimir los neurodos de escasa y constante actividad para todos los patrones, sino también, mediante la elección conveniente de la función de regularización, discriminar las regiones o zonas de activación que mayor impacto tienen sobre la performance de la Red.

II. Regularización

De acuerdo a lo mencionado, al error de entrenamiento (MSE) se le adiciona un término que considera las salidas de los neurodos ocultos. El algoritmo de BP realiza gradiente descendiente sobre la siguiente función de costo:

$$\frac{1}{2P}\sum_{j}^{P}\sum_{k}^{O}\left(t_{kj}-o_{kj}\right)^{2}+\mu\sum_{j}^{P}\sum_{l}^{H}R\left(o_{lj}^{2}\right) \qquad (1)$$

El primer término de esta función es la conocida suma sobre la diferencia cuadrática entre la respuesta deseada (t_{kj}) y el valor dado por la red (o_{kj}), para todos los patrones del conjunto de entrenamiento (CE) y todas las unidades de salida. P es el número de patrones y O la cantidad de neurodos de la capa de salida. Sin perdida de generalidad, en este trabajo se utiliza un solo nodo de salida. La suma en el segundo término se toma sobre un conjunto o subconjunto H de unidades ocultas. $\boldsymbol{R}$ es una función continua (ver Tabla 1) de la salida cuadrática del neurodo i-ésimo, μ es el parámetro de regularización.

El mínimo teórico de esta función se encuentra cuando las activaciones deseadas son iguales a las activaciones obtenidas para todas las unidades de salidas y para todos los patrones presentados y *además* cuando las unidades ocultas tienen una mínima contribución. Esto último significa lograr que el mayor número posible de unidades ocultas contribuyan mínimamente.

El algoritmo de BP siguiendo la notación standard[(1)], es adaptado para minimizar la función de costo dada por la Ec.(1). Ello involucra cambios en la variación de pesos, que pueden escribirse como:

$$\Delta w_{kl}=\alpha o_{l}\left(\delta_{k}^{\ error}+\mu\delta_{k}^{\ n}\right) \qquad (2)$$

Donde simplemente se ha adicionado un nuevo término que es:

$$\delta_{k}^{\ n}=2R'o_{k}f_{k}'(net_{k}) \qquad (3)$$

Aquí $\boldsymbol{R}'$ es la derivada de la función de regularización (ver Tabla 1), f' la derivada de la función logística, ***net*** es la conocida suma pesada de pesos y entradas, y w_{kl} es el peso de la conexión desde la unidad oculta l hacia la unidad de salida k.

Puesto que nos proponemos evaluar la metodología de regularización, aplicada en diversas zonas o regiones de la

activación de los neurodos ocultos, se eligen las diversas funciones de regularización de forma tal que actúen con mayor o menor intensidad de acuerdo al valor de las salidas. En la Tabla 1 se muestran las funciones elegidas y sus derivadas. Es la forma de estas últimas la que establece en cuales zonas de la activación su efecto será mayor (Ec.3 y Tabla 1). F1 actúa sobre todos los valores de las salidas por igual independientemente de su magnitud. F2 actúa decrecientemente para salidas cada vez mayores. F3 es una combinación de estas. F4 tiene un máximo en un valor intermedio entre las salidas grandes y pequeñas. Por último hemos considerado importante analizar una función exponencial, F5 que a diferencia de otros trabajos, actúa crecientemente sobre las salidas de mayor magnitud.

Funciones	F1	F2	F3	F4	F5
R	O^2	$\ln(1+o^2)$	$o^2+o^2/(1+o^2)$	$\ln[(1+o^2)\exp(1/1+o^2)]$	$\exp(o^2)$
$R'=dR/do^2$	1	$1/(1+o^2)$	$1+1/(1+o^2)^2$	$o^2/(1+o^2)^2$	$\exp(o^2)$

Tabla 1: Funciones de Regularización y sus derivadas

III. Simulaciones

A fin de analizar el procedimiento propuesto se han realizado simulaciones para evaluar la performance de aprendizaje y generalización.

Aprendizaje

El objetivo aquí es mostrar que la Red regularizada hace mejor uso de los recursos y es capaz de resolver el problema dado con un *número menor* de neurodos en la capa oculta. Se usan dos clásicos problemas de clasificación: O Exclusivo (Xor) y Simetría de 4 entradas[2].

Generalización

Para analizar habilidades de generalización, o sea la performance de la red para reconocer patrones nunca vistos, se diseñó un problema bidimensional (ver Fig.2) en que hay patrones de Tipo A y Tipo B. Las coordenadas x e y son las entradas (continuas) que definen ambos tipos de patrones. La red debe reconocer a que tipo de patrón pertenece el punto de coordenadas x e y que se le presenta. Se usó una red según se muestra en la Fig.1, utilizándose conjuntos independientes para entrenarla (CE) y para testear (CT).

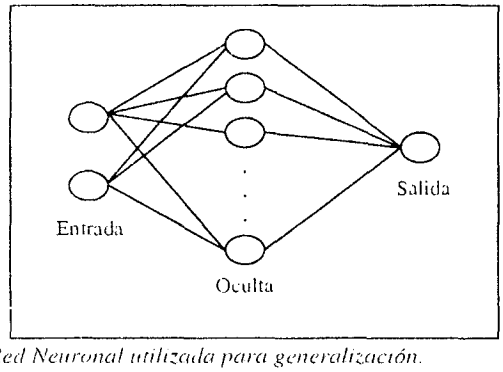

Fig.1: Red Neuronal utilizada para generalización.

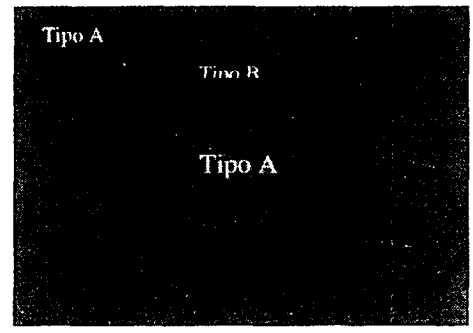

Fig.2: Problema bidimensional continuo para generalización.

IV. Análisis de los Resultados

Aprendizaje

Los resultados obtenidos para los problemas de clasificación sin y con regularización, se muestran en las Fig.3 y 4. Estas se construyeron de la siguiente manera: se inició al azar una red con 10 neurodos en la capa oculta, se entrenó hasta completar el aprendizaje y luego se ordenaron los neurodos ocultos de acuerdo a su activación y se eliminó uno por vez. Efectuada cada desactivación, se testeó la red remanente con el conjunto de entrenamiento (4 ejemplos para Xor y 16 para Simetría). Se realizaron 100 repeticiones del procedimiento y se contabilizó el número de redes que convergieron y el número de neurodos ocultos que tenia tal red para cada una de las funciones de regularización. Se visualiza así el *número efectivo de neurodos* que resuelven la tarea.

Los resultados que se muestran señalan la conveniencia de usar regularización y la superioridad de la función F3 para estos problemas. Concretamente para el caso de F3 se observa una convergencia de *ca* 100% aún cuando han sido desactivados 5 y 6 neurodos (quedan en la red 5 y 4 neurodos) para el caso del Xor y Simetría respectivamente. Es de notar el comportamiento de la función exponencial, F5, puesto que si bien el número de redes que convergieron es menor, este se mantiene constante hasta 2-3 neurodos ocultos, muy cercano al mínimo teórico para resolver estos problemas[1]. Se muestra que una función exponencial creciente como F5 da buenos resultados contrariamente a indicaciones en otros trabajos[2,3] de que salidas de pequeña magnitud son el candidato obvio para eliminación.

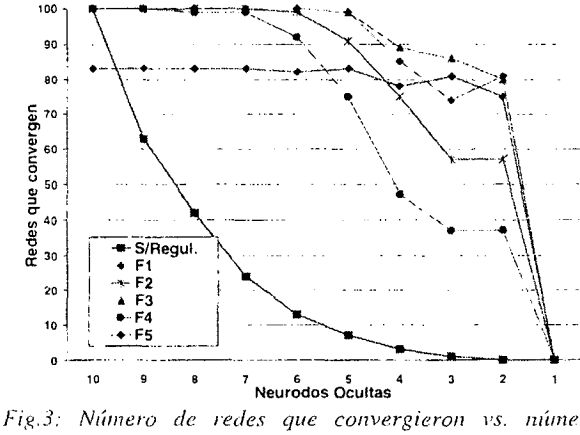

Fig.3: Número de redes que convergieron vs. número de neurodos en la capa oculta para Xor.

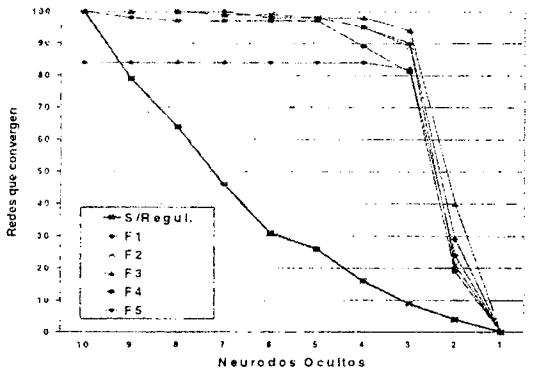

Fig.4: Idem Fig. 3 para Simetría.

Para estudiar generalización, en el problema continuo se realiza el análisis en dos etapas. *En la primera de ellas*, con el objetivo de determinar la o las funciones de regularización más aptas se procede similarmente a los problemas de clasificación. Se inician 30 redes con 10 neurodos ocultos cada una de ellas, una vez que se cumple el criterio de entrenamiento (85% de aciertos), se van eliminando neurodos de a uno por vez de acuerdo a su activación, usándose todas las funciones de regularización. La Fig.5 muestra los mejores resultados, obtenidos en este caso. El conjunto de entrenamiento de 50 patrones es también utilizado para testear las redes con menos de 10 neurodos ocultos.

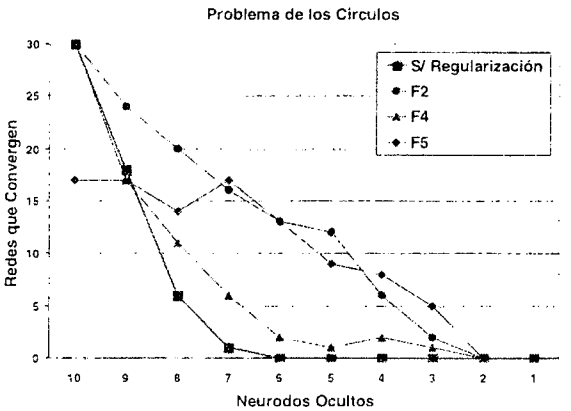

Fig.5: Porcentaje de aciertos vs. número de neurodos ocultos para diversas funciones de regularización.

Los resultados obtenidos muestran, la superioridad de la función F2, que da un decaimiento más suave con respecto al número de neurodos efectivamente usados. Es de notar aquí que F5 da resultados similares a los casos anteriores, la novedad es que F2 es la función que mejor se comporta. A primera vista esto pareciera contradictorio puesto que F2 y F5 actúan crecientemente sobre las salidas de magnitud pequeñas y grandes respectivamente. Sin embargo F2 converge mejor que F5 para 10, 9 y 8 neurodos, lo que determina su superioridad.

En la segunda etapa, se realiza una experiencia típica de generalización, en que se varia el número de patrones para entrenar y luego de cumplido el criterio de entrenamiento, se testea la red con los pesos así logrados. Los resultados se muestran en la Tabla 2 dando el número de aciertos en testeo (porcentaje) vs. el número de patrones utilizados en el entrenamiento. El Conjunto de Testeo (CT) fue de 5000 patrones tomados independientemente del CE. La estadística utilizada fue de 10 redes. En todos los casos de este trabajo se tomo μ = 0,01, con el que se lograron los mejores resultados.

Patrones para el Entrenamiento	Testeo			
	10 Neurodos s/Regul.	6 Neurodos Regul.F2	6 Neurodos Regul.F4	6 Neurodos Regul.F5
	%Aciertos	%Aciertos	%Aciertos	%Aciertos
100	84,07%	84,82%	84,26%	84,14%
50	81,52%	82,97%	83,00%	82,79%
40	76,84%	77,05%	76,66%	75,39%
30	75,26%	73,45%	73,85%	73,92%
20	59,01%	58,00%	59,54%	57,91%
10	57,95%	57,70%	57,70%	57,34%

Tabla 2: Se muestra el porcentaje de aciertos en testeo y el número de patrones para entrenar, para diversas funciones de regularización.

Como es evidente los resultados obtenidos con y sin regularización muestran el uso redundante de recursos. Debe notarse que el porcentaje de aciertos en testeo es en general igual o superior para las *redes regularizadas*, que tienen un *número menor* de neurodos ocultos.

V. Conclusiones

En este trabajo hemos mostrado:

1. La aplicación de un método de regularización a diversos problemas de aprendizaje y generalización.
2. Que aplicado a problemas de clasificación el método muestra una clara tendencia a generar la red de mínimo tamaño que resuelve el problema.
3. Que el método muestra también mejoras en la habilidad de generalización.
4. Que funciones no iguales en forma (F2, F3), pero cuyas derivadas tienen en común un decrecimiento hiperbólico con las salidas, se muestran las más aptas para ambos tipos de problemas.
5. Que es importante incluir la función exponencial como candidato en estos métodos de regularización.

Referencias

(1) Freeman, James A., Skapura David M. *Neural Networks. Algorithms, Applications, and Programming Techniques.* Addison-Wesley Publishing Company. Reading, Massachusetts. 1991.

(2) Hanson, S. J. Pratt, L. Y. En Advances in Neural Network Information Processing Systems, Vol. 1, pp.177-185. San Mateo, California: Morgan Kaufman. 1989.

(3) Chauvin, Y. En Advances in Neural Network Information Processing Systems, Vol. 1, pp.519-526. San Mateo, California: Morgan Kaufman. 1989.