Eliminacion de conexiones en redes neuronales recurrentes

J. A. Horas y C. P. Mankoc

Universidad Nacional de San Luis, Facultad de Cs. Fisico Matemáticas y Naturales,
Departamento de Fisica, Instituto de Matemática Aplicada San Luis.
Ejercito de los Andes 950, (5700) San Luis - Argentina
jhoras@unsl.edu.ar

Partiendo de una red tipo Hopfield totalmente conectada, se eliminan conexiones. El criterio para detener esta poda es la obtencion de similar performance. La performance se estudia y cuantifica determinando la capacidad y el tamaño de las cuencas de atraccion de las redes podadas. Se aplica tambien un procedimiento denominado de desaprendizaje que minimiza el numero de atractores espurios.

Starting from a full connected Hopfield network we delete synaptic weights. The stopping criterion for this pruning procedure is to obtain a similar performance. The latter is quantified by measuring the capacity and also the attraction basins' size of the pruning networks. We apply an unlearning procedure which minimize the spurious attractors number.

INTRODUCCION

Se ha probado que los cerebros de mamíferos durante la niñez y hasta la adolescencia reducen a menos de la mitad el número de sinapsis [1] . Esto se observa en distintas areas del cerebro tanto en animales como en humanos. Este costoso fenomeno aparentemente ineficiente hace suponer que existe una fuerte justificacion para este proceso. El estudio que aquí realizamos esta motivado por este notable fenómeno.

El objetivo de este trabajo es justamente, el de estudiar el comportamiento de redes neuronales al eliminar conexiones, sujeto a la obtención de similar performance. Se comparan procedimientos de poda y se somete a las redes a desaprendizaje. Para esta comparación se eligió utilizar el modelo de Hopfield [2] [3] , este es uno de los modelos de redes recurrentes de mas amplio uso, asimismo es interesante y muy utilizado para simular sistemas físicos como los vidrios de spin entre otros.

Por ultimo, y desde el punto de vista computacional, es posible obtener resultados completos e interesantes en poco tiempo de procesamiento.

DESCRIPCION DEL MODELO

La dinamica de la red esta dada por:

$$S_i(t+1) = sgn\left(\sum_j J_{ij} S_j(t)\right) \qquad (1)$$

En donde $S_i(t)$ denota el estado del i-ésimo nodo en el tiempo t, J_{ij} corresponde al peso que conecta el nodo j con el nodo i, los que se determinan a traves de la regla de Hebb:

$$J_{ij} = \frac{1}{N}\sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu \qquad (2)$$

N corresponde al número de nodos de la red, p es el numero de patrones a almacenar y ξ es uno de los patrones que se desea almacenar.

La red posee pesos simetricos $J_{ij} = J_{ji}$, lo que asegura la existencia de una función de Lyapunov [3] .

La memoria no funciona de manera perfecta, esto se debe a que cuando se graban numerosos patrones aparecen mínimos adicionales denominados estados espurios[2] .

Un método para reducir el numero de estados espurios y simultáneamente incrementar la capacidad de la red, aumentar el tamaño de las cuencas de atracción y permitir la operación con patrones correlacionados, es el de aplicar un procedimiento conocido como **desaprendizaje**[4] . Queremos determinar si al anterior listado, corresponde adicionar también un buen comportamiento en redes podadas. Para desaprender los estados espurios se inicializa la red en un patrón aleatorio, dejándose que relaje a un estado estable, una vez en este, se aplica la regla de Hebb "inversa" de tal forma que dicho estado sea "eliminado" de la memoria. Todo esto se repite un cierto numero de veces o pasos de desaprendizaje d.

La performance se puede medir por dos cantidades. Una es la **capacidad máxima**, que es el máximo numero de patrones que pueden almacenarse y ser recuperados establemente. Para el caso de las redes de Hopfield[5] ,$\alpha = \frac{p}{N} = 0.138$. La otra medida es el **tamaño de las cuencas de atracción**, esto es, cuan lejos de un patrón almacenado podemos inicializar la red siendo esta capaz de recuperarlo correctamente.

Una medida muy útil, de la similaridad entre dos patrones es el overlap, así si tenemos dos estados ζ_i y ξ_i con $i = 1...N$ el overlap entre ellos esta dado por

$$m = \frac{1}{N}\sum_j \zeta_j \xi_j \qquad (3)$$

esto da un numero entre 0 para estados ortogonales y 1 para estados idénticos.

En este trabajo, dado que la determinaciones de capacidad y tamaño de cuencas son muy costosas computacionalmente, proponemos utilizar la **estabilidad** de la red. Esta determinación[4] da buenas estimaciones tanto de la perfor-

mance como sobre el porcentaje de poda a utilizar, cantidad de pasos de desaprendizaje a aplicar y una buena comparación entre los diferentes métodos y procedimientos a usar. La estabilidad se puede expresar por:

$$\Delta_{\min} = \min \left\{ \xi_i^\mu \sum_j \xi_j^\mu J_{ij} \text{ con } i = 1..N,\ \mu = 1..p \right\} \quad (4)$$

Poda

Llamamos saliencia de un peso a una estimación de cuanto afectara a la performance la eliminación de dicho peso.

El primer método para eliminar conexiones es elegir cualquiera de los pesos aleatoriamente. Esto es obviamente un método basto de eliminar conexiones, ya que no todos los pesos influyen de la misma manera en el comportamiento de la red. Por lo tanto, si usamos este método, la saliencia de cada peso sera aleatoria.

Un método mas efectivo para determinar la eliminación de un peso, es utilizar su magnitud, esto se basa en que los pesos que tienen menor magnitud son los que menos afectan la sumatoria en la ec. (1), dado que los S_i toman valores de ± 1.

Si eliminamos un peso las conexiones de la red dejan de ser simétricas. Para evitar la asimetría y asegurar la convergencia a un minimo, cuando eliminamos un peso J_{kl}, también eliminamos el reciproco J_{lk}.

El procedimiento de poda puede generar nuevos estados espurios. Para eliminar estos posibles estados, tras cada poda se efectúa un procedimiento de desaprendizaje. Este podría producir alguna degradación de la información almacenada en la red, dado que solo se cuenta con una estimación del numero de pasos de desaprendizaje óptimo. En la eventualidad de que esto ocurra se efectúa una etapa de "reaprendizaje", en la cual se modifican los pesos *que no han sido podados,* de acuerdo a la ecuacion:

$$J_{ij}^{(n+1)} = J_{ij}^{(n)} + \frac{\nu}{N} \sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu \quad (5)$$

SIMULACIONES

Dado que este problema es extremadamente complejo y no lineal, se efectuaron simulaciones numéricas para mostrar la efectividad de los métodos de poda.

Las simulaciones se realizaron sobre redes de 200 nodos, esta es una red suficientemente grande como para obtener resultados interesantes pero suficientemente pequeña como para lograr estos en tiempos cortos. La actualización de las redes fue asincrona, los patrones almacenados fueron aleatorios, con igual probabilidad de ser ± 1 cada uno de sus bits.

Como nos interesa el comportamiento de la red a medida que se eliminan pesos, procedemos según:

1. Calcule los pesos J_{ij} de la red usando la ec. (2).
2. Efectúe d_i pasos de desaprendizaje inicial .
3. Elimine pesos usando el método que corresponda.
4. Efectúe el reaprendizaje.
5. Efectúe un numero de pasos de desaprendizaje d.
6. Efectúe la medición de la estabilidad de la red.
7. Si aun puede podar, repita desde el paso 3

El paso 2 se aplica a fin de comenzar el procedimiento de poda con una red a la que se le eliminaron los estados espurios mas importantes. Por simplicidad, el numero de pasos de desaprendizaje d_i y d fueron iguales.

En la tabla 1 se muestran los diferentes métodos que fueron aplicados, indicando los items del esquema anterior que se efectuaron (tilde ✓) o no (cruz ×) como así también los pasos de desaprendizaje aplicados.

TABLA 1: ITEMS Y PASOS DE DESAPRENDIZAJE LLEVADOS A CABO EN CADA METODO.

Metodo de Poda	Items						Pasos
	1	2	3	4	5	6	
Azar	✓	×	✓	×	×	✓	–
Azar con desaprendizaje	✓	✓	✓	×	✓	✓	140
Azar con reaprendizaje	✓	✓	✓	✓	✓	✓	240
Minimos	✓	×	✓	×	×	✓	–
Minimos con desaprendizaje	✓	✓	✓	×	✓	✓	60
Minimos con reaprendizaje	✓	✓	✓	✓	✓	✓	90

En la fig. 1 se muestran las curvas de estabilidad versus el porcentaje de la red que ha sido podado. Se realizaron múltiples pruebas, se muestra el mejor resultado obtenido para cada uno de los métodos de poda, con el numero de pasos de desaprendizaje dado en la tabla 1.

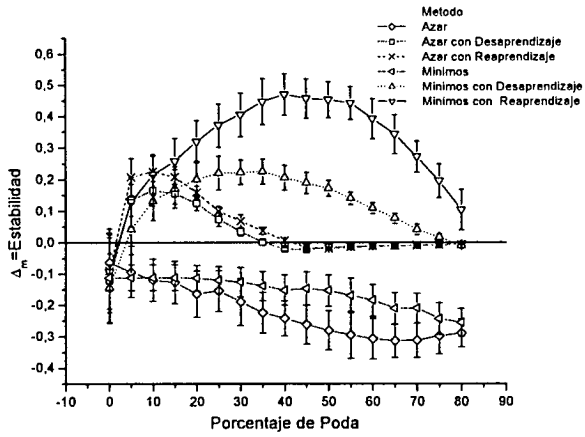

Figura 1: Se muestra la estabilidad versus el porcentaje de pesos eliminados (ver texto).

Esta grafica permite comparar el comportamiento de redes de igual numero de pesos que han sido podadas por metodos diferentes. Las curvas que se muestran tienen una carga $\alpha = 0.2$, un parametro de desaprendizaje $\varepsilon = (-)0.001$ y para las curvas con reaprendizaje un parametro de reaprendizaje $\nu = 0.05$; cada uno de los puntos corresponden a un promedio sobre 25 redes y las barras de error indican la desviacion standard de cada una de las medidas.

Como se observa en la fig 1 la poda por mínimos (con desaprendizaje y con reaprendizaje) presenta ventaja con respecto a los otros métodos en el valor máximo que alcanza la estabilidad. También es de notar (ver tabla 1) que el método de mínimos requiere un menor número de pasos de desaprendizaje entre poda y poda para obtener la mejor curva. Esto es atribuible al hecho que la elección aleatoria de los pesos a podar elimina con igual probabilidad tanto pesos de poca relevancia como pesos importantes, lo que

produce mas cantidad de estados espurios y se requiere un mayor numero de pasos de desaprendizaje para eliminarlos. Asimismo se observa que es necesario también hacer un mayor numero de pasos de desaprendizaje en las redes en las que se le efectúa el reaprendizaje, ya que este agrega nuevamente estados espurios.

La estabilidad, como se ha dicho, da una buena estimación de la performance, para explorar con mas fineza zonas relevantes, damos medidas de capacidad y de cuencas en esas regiones.

Capacidad máxima

Para observar el comportamiento de la capacidad de la red al eliminar pesos, en las figuras 2 y 3 se muestran las curvas de capacidad para los diferentes métodos utilizados y para los porcentajes de poda que se indican.

El procedimiento general para obtener estas gráficas es similar al usado para obtener la fig (1), con la diferencia que en el paso 6 se efectúa una medida del overlap tras la relajación cuando se inicializa la red desde uno de los patrones, esto se efectúa una gran cantidad de veces para determinar el valor promedio de m.

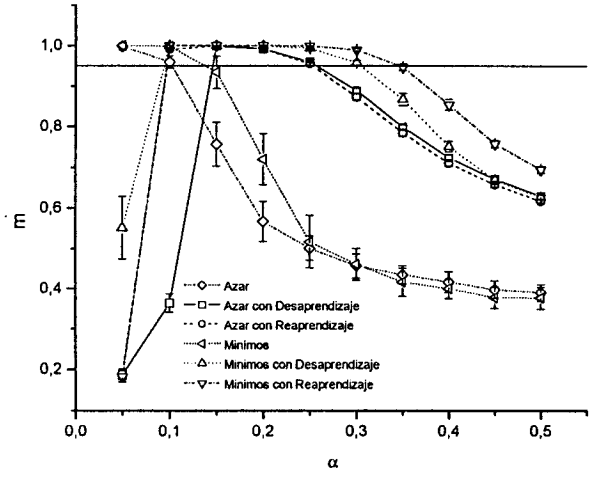

Figura 2: Overlap promedio (m) versus carga (α) para los diferentes métodos de poda. Poda 45%

En las gráficas de capacidad donde se aplicó desaprendizaje, se observa que para muy poca carga el overlap es allí extremadamente bajo, esto se debe a que el número de pasos de desaprendizaje aplicados fueron excesivos para el pequeño número de patrones almacenados.

En la tabla 2 se muestran los valores de la capacidad máxima para los distintos métodos de poda y para un porcentaje creciente de esta ultima. Se observa que los resultados mejoran con la aplicación del procedimiento de desaprendizaje en todos los casos en que se aplica, y que la poda por mínimos supera al método de poda al azar.

Tamaño de las cuencas de atracción

La otra medida para determinar la performance esta dada por el tamaño de las cuencas de atracción. Una buena estimación de dicho tamaño puede obtenerse realizando las típicas curvas que comparan overlap inicial con overlap final[3] .

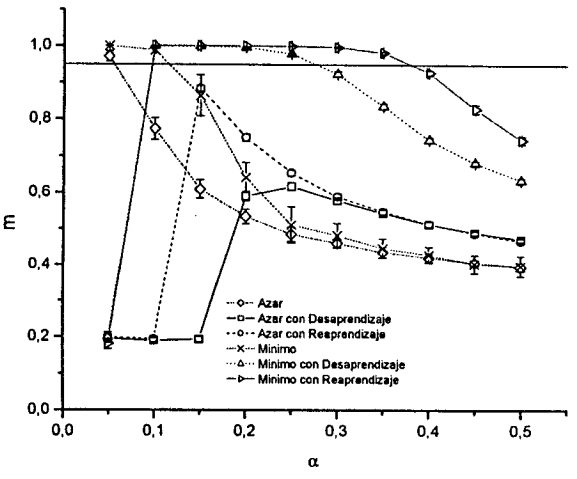

Figura 3: Overlap promedio (m) versus carga (α) para los diferentes métodos de poda. Poda 75%

TABLA 2: CAPACIDAD DE LA RED PODADA PARA LOS DISTINTOS METODOS UTILIZADOS.

Capacidad Maxima				
Metodos	Porcentaje de poda:			
	15%	45%	55%	75%
Azar	0,129	0,103	0,081	0,054
Azar con Desaprendizaje	0,228	0,257	0,216	-
Azar con Reaprendizaje	0,256	0,254	0,211	-
Minimos	0,129	0,135	0,127	0.114
Minimos con Desaprendizaje	0,261	0,305	0,311	0,275
Minimos con Reaprendizaje	0,285	0,345	0,355	0,379

En las figs. 4 y 5 se observan las gráficas de overlap final promedio versus el overlap inicial para 45% y 75% de pesos podados en la red para cada uno de los métodos.

Los insets en las figs.4 y 5 muestran una ampliación de la zona de interes ($m \geq 0.95$)

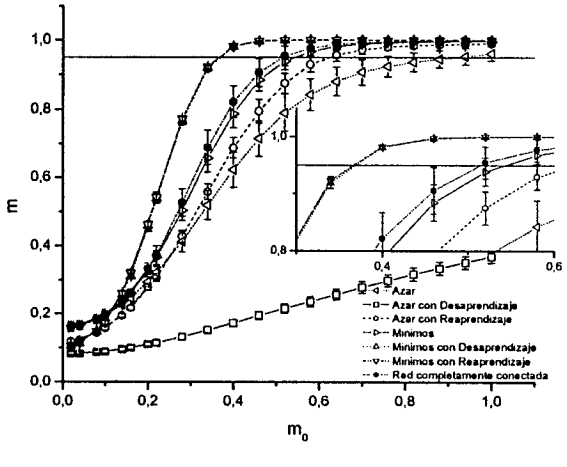

Figura 4: Overlap final (m) versus overlap inicial (m_0) para los distintos metodos de poda. 45% de pesos podados.

En la tabla 3 se muestra el overlap inicial para los que las graficas cortan la linea $m = 0.95$, para cada uno de los metodos y los respectivos porcentajes de poda.

Se observa también que el método de poda por mínimos es superior al metodo de poda al azar, especialmente

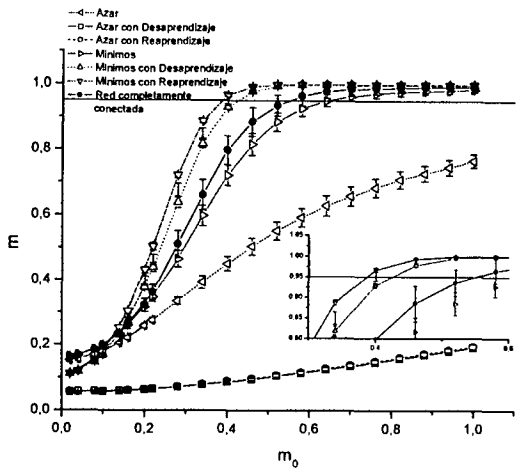

Figura 5: Overlap final (m) versus overlap inicial (m_0) para los distintos metodos de poda. 75% de pesos podados.

para altos valores de poda. La aplicación de desaprendizaje se muestra importante sobre el comportamiento de la red para todos los métodos utilizados. Pero el uso del procedimiento de reaprendizaje, en particular para el caso de poda por mínimos, no tiene un efecto notable. Observar únicamente esto, introduce la duda de si estamos durante el reaprendizaje solo agregando información a la red que luego eliminamos en el desaprendizaje. Por ello y para observar como evolucionan los pesos a medida que se efectúa la poda se realizaron histogramas de los valores de los pesos de la red mientras se efectúa la poda.

TABLA 3: SE MUESTRAN EL OVERLAP INICIAL PARA EL PUNTO DE CORTE $m = 0.95$

Tamaño de Cuencas				
Metodos	Porcentaje de poda:			
	15%	45%	55%	75%
Azar	0,57	0,92	-	-
Azar con Desaprendizaje	0.38	-	-	-
Azar con Reaprendizaje	0,38	0,62	-	-
Minimos	0,53	0,53	0,56	0,64
Minimos con Desaprendizaje	0,39	0,37	0,38	0,42
Minimos con Reaprendizaje	0,38	0,37	0,37	0,39

Histogramas del valor de los pesos.

Se muestran, en las figs. 6 y 7, los histogramas del valor de los pesos en la red para diversos porcentajes de poda por mínimos con desaprendizaje y con reaprendizaje.

En la fig 6 se observa que los pesos podados son efectivamente los de menor valor, produciendose la correspondiente separación entre los valores positivos y negativos de los pesos a medida que se produce la poda.

A diferencia de lo anterior, en la fig. 7 la aplicación del procedimiento de reaprendizaje en poda por mínimos produce un ensanchamiento de la base del histograma. Esto permite deducir que efectivamente estamos alterando los pesos de la red y no quitando en el desaprendizaje lo que agregamos durante el reaprendizaje.

CONCLUSIONES

De este trabajo se puede concluir:

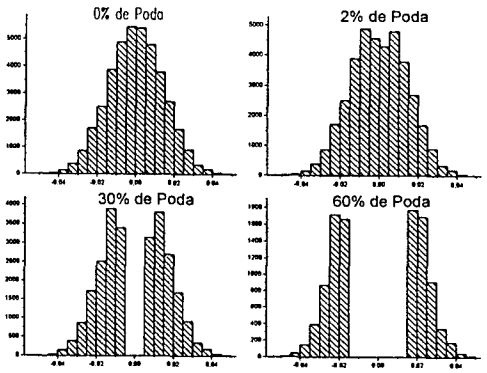

Figura 6: Histogramas donde se aplicó solo desaprendizaje.

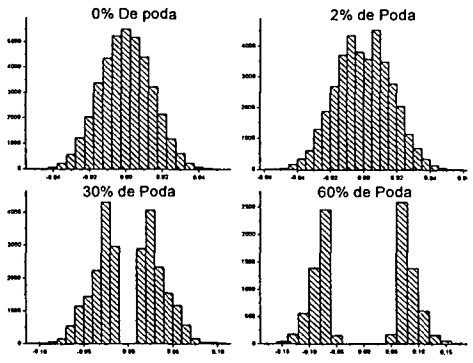

Figura 7: Histogramas donde se ha aplicado desaprendizaje y tambien reaprendizaje.

1. La estabilidad se ha mostrado apta como un buen estimador de la performance, y permite también establecer, entre otros, los pasos de desaprendizaje correctos para no degradar la memoria.
2. Las redes podadas en porcentajes importantes mantienen o aún acrecientan la performance de las redes originales, mas densas en las conexiones.
3. Al desaprendizaje y sus múltiples ventajas hay que adicionar su buen comportamiento en el procedimiento de poda.
4. El método de poda que se mostró mas eficiente es el de mínimos con desaprendizaje.
5. El procedimiento de reaprendizaje no es absolutamente necesario.

Bibliografia

[1] G. M. Innocenti. Neurosci,18:397-402, 1995.

[2] J. Hertz, A. Krogh y R. G. Palmer. "*Introduction to the theory of neural computation*". Addison-Wesley. 1994.

[3] E. Domany, J.L. Van Hemmen y K. Schulten "*Models of neural networks*". Springer Verlag, 1991.

[4] J. A. Horas y P. M. Passinetti, J. Phys. A: Math. Gen. 31 (1998) L463-L471

[5] D. Amit, H. Gutfreund y H. Sompolinsky. *Physical Review A* 35. 2293-303. 1987.