Learning experiments with High Order Boltzmann Machines

keywords:Neural Networks, Boltzmann Machines, High Order Networks

Abstract: This paper presents the results obtained applying High Order Boltzmann Machines without hidden units to a several well known problems that represent quite diverse learning paradigms. The learning algorithm of Boltzmann Machines easily generalises for machines discrete units (whose state spaces are discrete multivalued sets) and continuous units. The learning algorithm remains valid for continuous units not normalised in the [0,1] interval. Besides, the absence of hidden units allows the estimation of the mean activation levels (activation probabilities in the binary case) without using simulated annealing. This provides a substantial acceleration of the learning process, and a qualitative improvement of the results. The presentation, description and discussion of the experiments follows the historical order of our works.

Experimentos de aprendizaje con Máquinas de Boltzmann de alto orden

M. Graña, A. D´Anjou, F.X. Albizuri, J.A. Lozano, P. Larrañaga, Y. Yurramendi, M. Hernandez, J.L. Jimenez, F.J. Torrealdea, M. Poza, A.I Gonzalez
Dept. CCIA, UPV/EHU+
Apartado 649, 20080 San Sebastián
e-mail: ccpgrrom@si.ehu.es

Resumen: En el presente trabajo se presentan los resultados obtenidos en la aplicación de las Máquinas de Boltzmann de alto orden sin unidades ocultas a un conjunto de problemas que representan paradigmas muy diferenciados de problemas de aprendizaje. El algoritmo de aprendizaje de las Máquinas de Boltzmann se generaliza para Máquinas con unidades discretas (cuyos estados son valores en un conjunto discreto no necesariamente binario) y continuas. El algoritmo de aprendizaje mantiene su validez para unidades continuas con espacios de estado no normalizados al intervalo [0,1]. Se observa, además, que la ausencia de unidades ocultas permite prescindir del enfriamiento estadístico (simulated annealing) en la estimación de los niveles de activación promedio (probabilidades de activación en el caso binario), lo cual permite una aceleración substancial del proceso de aprendizaje, y una mejora cualitativa de los resultados. La descripción y presentación de los experimentos sigue el orden histórico de nuestros trabajos.

Palabras clave: Redes neuronales, Máquinas de Boltzmann, Redes de alto orden

0 Introducción

La Máquina de Boltzmann es una de las primeras proposiciones de arquitecturas de redes neuronales [1, 2]. Ha sido relegada en su aplicación práctica (comercial) debido a lo costoso del algoritmo de aprendizaje y a la dificultad de ajustar ciertos parámetros. Por ejemplo, resulta difícil determinar las temperaturas finales en los procesos de enfriamiento estadístico. Su buen ajuste es crítico para la correcta toma de estadísticas de los niveles de activación de las conexiones. Nuestros trabajos siguen las convenciones adoptadas por Aarts [1], que considera la Máquina de Boltzmann como un maximizador de la función de consenso, en el que los estados de las unidades (en el caso binario) son 0 y 1. En este artículo se retoma esta arquitectura con la pretensión de demostrar que su utilización práctica es factible. Esto proviene de dos condiciones sobre la topología: la ausencia de unidades ocultas y la utilización de conexiones de alto orden. (El orden de una conexión es el número de unidades que conecta. Las

+ Trabajo realizado dentro de los proyectos PGV9220, financiado por el Dept. Educación, Univ. e Inv. of the Gobierno Vasco, y UPV140.226-EA166/94, financiado por la UPV/EHU

arquitecturas neuronales "convencionales" utilizan sólo conexiones de orden 2. Llamamos conexiones de alto orden a conexiones de orden mayor que dos).

En realidad ambas condiciones están relacionadas, puesto que tanto las (conexiones hacia) unidades ocultas como las conexiones de alto orden capturan correlaciones de alto orden de los datos. El empleo de conexiones de alto orden, por tanto, permite evitar las unidades ocultas. Además, la Máquina de Boltzmann sin unidades ocultas tiene una interesante propiedad: la distancia de Kullback-Leibler entre la distribución de probabilidad del problema (de los datos) y la modelada por la Máquina de Boltzmann es, en general, convexa, y tiene un único mínimo global [4]. Esto permite dos simplificaciones del algoritmo de aprendizaje de importancia práctica fundamental:

- Los pesos iniciales de las conexiones pueden ser fijados arbitrariamente (en este artículo siempre a cero) y no es necesario explorar otras condiciones iniciales para asegurarse de que no existen mínimos mejores.

- La toma de estadísticas puede simplificarse, incluso llegando a obviar la necesidad de realizar enfriamiento estadístico (simulated annealing), en problemas de clasificación.

El aprendizaje en las Máquinas de Boltzmann de alto orden sin unidades ocultas es, por tanto, mucho más robusto y varios ordenes de magnitud más rápido que en el caso de las Máquinas de Boltzmann convencionales con unidades ocultas, (prácticamente las únicas consideradas hasta ahora en la literatura). La única referencia, previa a nuestros trabajos, que encontramos a las Máquinas de Boltzmann de alto orden es una nota de Sejnowski [20] que no hace sino apuntar una definición. Otras referencias a topologías neuronales con conexiones de alto orden son del estilo de [16] donde las conexiones de orden 3 sirven para obtener clasificadores invariantes a la traslación, escalado y rotación de los patrones. Pero los pesos de dichas conexiones no son "aprendidos" sino calculados *a priori* en función de las características geométricas del problema. Merece especial mención el trabajo de Pinkas [17], en el que relaciona las unidades ocultas y las conexiones de alto orden (unidades sigma-pi en su trabajo) en el contexto del problema de la satisfacibilidad en la lógica de proposiciones. Pinkas proporciona algoritmos que permiten obtener la topología de orden dos con unidades ocultas equivalente a una topología de alto orden dada, y viceversa. Parte substancial de la originalidad de nuestro trabajo reside en que aplicamos el algoritmo de aprendizaje a todas las conexiones, y es el proceso de aprendizaje el que determina la significación de las correlaciones de alto orden modeladas por las conexiones de alto orden.

Queremos prevenir al lector de que las descripciones topológicas que aparecerán son un tanto engorrosas. Están realizadas en notación de conjuntos y no existen representaciones gráficas. La razón para ello es doble. Por un lado no conocemos de convenciones gráficas aceptables para la representación de hipergrafos, y los intentos que hemos hecho conducen a diagramas muy difíciles de interpretar (y de dibujar), que no aclaran gran cosa. La segunda es que la forma más apropiada de definir las conexiones de alto orden es como conjuntos de unidades, por lo que el lenguaje más apropiado para la descripción topológica resulta ser el de la teoría de conjuntos.

Es posible que los problemas de implementación (los pesos se representan mediante un tensor en lugar de una matriz, las conexiones forman un hipergrafo en lugar de un grafo) de arquitecturas de alto orden hayan prevenido la extensión de su aplicación. Los

programas que realizan los experimentos (Actualmente escritos en ADA. Estamos ultimando una versión más eficiente en C++) pueden ser obtenidos mediante FTP anónimo en el nodo ftp.sc.ehu.es en el directorio pub/unix/hobm.

Otra aportación de este trabajo es la generalización del algoritmo de aprendizaje para Máquinas de Boltzmann con unidades no binarias, en particular para unidades cuyos estados pueden ser números enteros y para unidades con estados continuos. Es importante notar, como se demuestra experimentalmente sobre el problema de reconocimiento de vocales, que los estados no necesitan estar normalizados, esto es, el algoritmo de aprendizaje funciona para unidades continuas con espacios de estado distintos al intervalo [0,1]. Hasta este momento (véase por ejemplo [12, 15, 14]), las únicas arquitecturas neuronales que admitían unidades con estados no normalizados eran las arquitecturas competitivas, basadas en la búsqueda del vecino más cercano en términos, usualmente, de la distancia euclidea.

La organización del artículo refleja la secuencia de nuestros trabajos y hemos considerado que es de interés mostrar esta evolución, en la medida en que ilustra las motivaciones que impulsaron y los puntos de inflexión en los que se imponía una variación sobre lo aceptado como norma hasta el momento. El punto de partida fue la asunción como arquitectura básica de las Máquinas de Boltzmann de alto orden sin unidades ocultas y estados binarios. El algoritmo de aprendizaje inicialmente aplicado era relativamente clásico en la literatura (con enfriamiento hasta la temperatura de toma de estadísticas, fijando los inputs en la fase libre, tomando estadísticas sobre todos los patrones de entrenamiento (aprendizaje off-line) antes de proceder a la adaptación de los pesos sumando o restando 1).

Para los problemas de MONKS's es posible deducir una topología *a priori*, basándose en la interpretación de las conexiones de alto orden como operadores AND generalizados. Esta topología *a priori* sirvió para verificar la aplicabilidad de las Máquinas de Boltzmann a estos problemas, y permitió plantear el aprendizaje sobre topologías generales como aproximaciones a la búsqueda de dicha topología *a priori*. En este contexto se probaron dos tipos de algoritmos de poda, que no ofrecieron resultados alentadores. La dificultad de resolver el segundo de los problemas de MONK's, partiendo de topologías generales, nos llevó a considerar unidades con estados discretos no necesariamente binarios, y a generalizar consecuentemente el algoritmo de aprendizaje. También, en este caso es posible diseñar una máquina *a priori* para el segundo de los problemas MONK's, y también se probaron los algoritmos de poda, con los mismos pobres resultados.

La aplicación al reconocimiento de señales sonar nos hizo patente la dificultad de ajustar la temperatura de toma de estadísticas en el caso de unidades discretas con rangos muy grandes, lo que nos llevó naturalmente a considerar unidades con estados continuos, relajando la toma de estadísticas (a temperatura fija, sin enfriamiento). Por último, la aplicación al problema de reconocimiento de vocales nos permitió verificar experimentalmente la validez del algoritmo de aprendizaje para estados continuos no normalizados al intervalo [0,1]. Buscando acelerar los experimentos de reconocimiento de vocales, realizamos el aprendizaje obviando el enfriamiento estadístico con el resultado no esperado de un incremento substancial en la calidad del aprendizaje. La modificación de los pesos utilizando el gradiente estricto y un término de momento nos proporcionó, para este problema, resultados comparables a los de referencia. En ninguno de estos dos últimos problemas se aplicaron algoritmos de poda al aprendizaje.

A los lectores familiarizados con las Máquinas de Boltzmann, la afirmación de que se puede obviar el enfriamiento estadístico en el proceso de aprendizaje les resultará, sin duda, chocante en primera instancia. Queremos enfatizar, en primer lugar los beneficios. El aprendizaje se hace mucho más rápido y exacto. En segundo lugar, es necesario delinear las condiciones en que esto es posible:

> -En máquinas ***sin*** unidades ocultas. Cuando no existen unidades ocultas, el enfriamiento estadístico no es necesario para estimar la distribución de probabilidad estacionaria de sus estados.
> - En problemas de ***clasificación*** en los que deseamos estrictamente asociar una clase a un patrón input. En este caso tiene sentido la variante del algoritmo básico de aprendizaje en el que las únicas unidades libres en la fase libre del aprendizaje son las unidades output.
> -En problemas donde cada unidad output ***identifica*** una clase. En este caso, la búsqueda de la respuesta de la red (la clase más probable) es una búsqueda (de complejidad lineal en el número de unidades output) de la unidad output con máximo consenso.

En cualquier otro caso, el enfriamiento estadístico es imprescindible para poder estimar los niveles de activación promedio de las conexiones. Cuando las condiciones anteriores se cumplen, la aplicación del enfriamiento estadístico no hace sino introducir ruido innecesario y cargar computacionalmente el proceso de ajuste de los pesos.

Bajo el punto de vista de Máquinas de Boltzmann como un mecanismo de cálculo distribuido, obviar el enfriamiento estadístico ataca la idea misma del cálculo distribuido (donde cada unidad realizaría una búsqueda local asíncronamente). Sin embargo, hemos de notar que la posibilidad de realizar realmente dicho proceso absolutamente en paralelo tiene fuertes limitaciones [22, 23] sobre todo debido a la contribución del propio cálculo en paralelo a la función de seudo-consenso que es la realmente optimizada. Dicha contribución aumenta con el grado de paralelismo, hasta el extremo de hacer negligible a la verdadera función de consenso. A la vista de estos resultados, creemos que mantener por principio la asociación entre enfriamiento estadístico y Máquinas de Boltzmann es innecesario.

A continuación presentamos brevemente los problemas sobre los que se han realizado los experimentos, resumiendo nuestra aproximación a ellos. En la sección 1 se introducen los problemas de MONK's, en la sección 2 se exponen los resultados de la aplicación de las Máquinas de Boltzmann de alto orden con unidades binarias a estos problemas, en la sección 3 se presenta la extensión del algoritmo de aprendizaje al caso de Máquinas de Boltzmann con unidades discretas (no binarias), en la sección 4 se exponen los resultados de las MB discretas sobre los problemas de MONK's. En las sección 5 se introduce el problema del reconocimiento de señales sonar y los resultados con máquinas discretas y continuas. En la sección 6 se presentan los resultados de las máquinas continuas sobre un problema de reconocimiento de vocales en inglés. Finalmente, en la sección 7 se ofrecen conclusiones y lineas de trabajo futuro. En el apéndice se ofrece una derivación del algoritmo de aprendizaje para máquinas con unidades discretas.

0.1 Los problemas test

Los problemas de aprendizaje sobre los que se han aplicado las Máquinas de Botzmann de alto orden tienen las siguientes características en común.

1- Los datos son de dominio público, accesibles vía ftp.
2- Existen referencias de experimentos de aprendizaje sobre estos datos utilizando otras técnicas. Estas referencias permiten asegurar de forma relativamente objetiva la calidad de los resultados obtenidos con las Máquinas de Boltzmann de alto orden.
3- El método experimental está definido, en general mediante la separación de los datos en datos de entrenamiento y datos de test.
4- Son problemas de clasificación. El output de la red es siempre un vector binario, donde la clase está indicada por el único componente de valor 1. Esto reduce la búsqueda de la respuesta de la red a una búsqueda de complejidad lineal del máximo. (Recuérdese que no consideramos unidades ocultas y el input está fijado en la fase libre)

Los problemas de MONK's [21] se utilizaron como benchmark para la evaluación de un conjunto amplio de algoritmos de aprendizaje (Machine Learning), en su mayor parte algoritmos de construcción de árboles de clasificación. Entre otros, se probaron también dos algoritmos de redes neuronales: Backpropagation y Cascade Correlation. El presente trabajo puede considerarse, en parte, como una extensión de la evaluación realizada en [21] para incluir las Máquinas de Boltzmann de alto orden, con unidades binarias y no binarias.

Parte de los resultados que hemos obtenido para los problemas de MONK's han sido referidos en [11]. Nuestro punto de partida en la aplicación de las Máquinas de Boltzmann de alto orden es doble. Por un lado la demostración en [3] de que las Máquinas de Boltzmann de alto orden con unidades binarias pueden entrenarse con el mismo algoritmo que las convencionales y son capaces de ajustar cualquier distribución de probabilidad binaria . Por otro lado, la interpretación de las conexiones (para MB binarias en el espacio $\{0,1\}^N$) como cláusulas en forma conjuntiva. En [6,7] hemos explorado con buenos resultados esta posibilidad. Esta interpretación permite, a partir de la definición lógica de los problemas de MONK's, definir las topologías *a priori* que mejor se ajustan a cada uno de los problemas. Obviamente, dicha definición no es usualmente conocida, pero la existencia de la topología ideal permite testar el algoritmo de adaptación de los pesos sin interferencias topológicas y evaluar la capacidad de dicho algoritmo de descubrir la topología ideal. La dificultad del segundo de los problemas motivo la consideración de unidades no binarias. Para las máquinas no binarias la interpretación lógica de las conexiones deja de ser válida, y la tarea de definir una topología ideal viene a ser idéntica a la no trivial de transcribir el problema de aprendizaje en uno de optimización combinatoria.

El problema del reconocimiento de señales sonar es el estudiado en [10] utilizando redes entrenadas con el algoritmo de retropropagación (backpropagation) del gradiente. Es una de las referencias clásicas en la literatura de redes neuronales, puesto que es uno de los primeros trabajos en los que se hace un estudio de la calidad del aprendizaje en función del número de unidades ocultas (de la topología). Nuestra aproximación inicial a este problema fue la aplicación de las máquinas con unidades discretas, pero

encontramos fuertes dificultades en la estimación del parámetro temperatura, debidas a los rangos de valores que podían tomar los estados de las unidades y a la propia definición multiplicativa de la función de consenso. Esto nos indujo a considerar unidades con estados continuos para representar las variables input.

El último de los casos estudiados viene dado por un conjunto de datos sobre reconocimiento de vocales, que fueron utilizados para la realización de sendas tesis doctorales [8, 18, 19]. La segunda de estas tesis consistió, al menos en parte, en la comparación empírica de diversas técnicas conexionistas. El problema consiste en reconocer las vocales inglesas usando diez características obtenidas a partir la señal digitalizada. Los detalles de la captura de datos y del cálculo de las características pueden encontrarse en las tesis mencionadas y en la propia descripción del problema que se obtiene vía ftp de la base de datos de la CMU. Nuestro interés en este problema no consiste (por el momento) en proponer las Máquinas de Boltzmann como alternativa para el reconocimiento de voz (problema arduo donde los haya), sino como test de su capacidad de aprendizaje. Como en los casos anteriores, los resultados obtenidos mediante otras técnicas han servido como benchmark para las Máquinas de Boltzmann.

1 Definición de los problemas de MONK's

Los problemas de MONK's están definidos sobre un dominio de robots artificiales, donde un robot se describe mediante seis atributos o variables que toman valores discretos:

x_1 : head_shape {round, square, octagon}
x_2 : body_shape {round, square, octagon}
x_3 : is_smiling {yes, no}
x_4 : holding {sword, balloon, flag}
x_5 : jacket_color {red, yellow, green, blue}
x_6 : has_tie {yes, no}

Los conjuntos de datos de entrenamiento y test que definen cada uno de los problemas son de dominio público y se obtuvieron vía ftp del directorio "pub/neuroprose" del nodo "archive.cis.ohio-state.edu". A continuación reproducimos la definición formal de cada problema:

Problema M_1 definido por la relación

(head_shape = body_shape) or (jacket_color = red)

El conjunto de entrenamiento está formado por 124 ejemplos, elegidos aleatoriamente entre los 432 posibles patrones.

Problema M_2 definido por la relación

Exactamente dos de los seis atributos tienen su primer valor.

El conjunto de entrenamiento consta de 169 ejemplos, también elegidos aleatoriamente.

Problema M_3 definido por la relación

(jacket_color is green and holding a sword) or
(jacket_color is not blue and body_shape is not octagon).

El conjunto de entrenamiento consta de 122 ejemplos, elegidos aleatoriamente y, contrariamente a los problemas anteriores, con un 5% de clasificaciones erróneas ("ruido").

M_1 es una forma normal disyuntiva estándar, por lo que se supone que es fácil de aprender por cualquier algoritmo de aprendizaje simbólico (i. e. AQ, ID3). Por el contrario, M_2 es similar a los problemas de paridad, combinando los diferentes atributos de forma tal que dificulta su descripción en DNF o CNF. M_3 se utiliza para evaluar los diferentes algoritmos en presencia de ruido.

La metodología experimental seguida para los problemas de MONK's es la propuesta en [21]: cada problema está definido por dos conjuntos de datos: los datos de entrenamiento y los datos de test. La red se entrena con los primeros y se prueba con los segundos. Siguiendo el formato propuesto, los resultados son los porcentajes de aciertos sobre el conjunto test directamente. Se asume que todos los algoritmos son capaces de fijar completamente el conjunto de entrenamiento.

2 Resultados de las Máquinas de Boltzmann binarias de alto orden aplicadas a los problemas de MONK's

La notación que empleamos es la siguiente: Una máquina de Boltzmann viene descrita por la tripleta (U,L,W), siendo U el conjunto de las unidades y L el de las conexiones. Una conexión λ L es un subconjunto de U. El orden de una conexión es su cardinalidad y el orden de la máquina de Boltzmann es el la conexión de mayor orden. W es el conjunto de pesos asociados a las conexiones, y puede formularse como una aplicación W:L∅R. La forma de la función de consenso es $C(k) = \sum_{\lambda} \omega_\lambda \prod_{u\in\lambda} k(u)$ donde k(u) es el estado de la unidad u en la configuración global k. Asumimos que la mecánica de la Máquina de Boltzmann consiste en optimizar la función de consenso [1], y el aprendizaje trata de asociar máximos locales idénticos a las configuraciones que se desea aprender.

En la aplicación a los problemas MONK's el conjunto de unidades binarias utilizado en los tres problemas tiene 16 unidades y se construye de la siguiente manera:

$$U_1 = \left\{u_{ij} \middle| i \in \{1,2,4\}, j \in \{1..3\}\right\}$$

$$U_2 = \left\{u_{ij} \middle| i \in \{3,6\}, j = 1\right\}$$

$$U_3 = \left\{u_{ij} \middle| i = 5, j \in \{1..4\}\right\}$$

$$U^{16} = U_1 \cup U_2 \cup U_3 \cup \{u_o\}$$

En esta representación, para variables cuyos rangos son no binarios, la unidad u_{ij} toma el estado 1 cuando la variable x_i toma el j-esimo valor de su conjunto rango. Para representar las variables x_3 y x_6, de rango binario, se emplea una única unidad. La unidad u_o representa el output de la clasificación.

Basándonos en la interpretación de las conexiones como cláusulas AND, se puede deducir para cada uno de los problemas una topología *a priori*. A continuación expresamos los conjunto de conexiones que constituyen la topología *a priori* para cada uno de los problemas, en la representación con 16 unidades dada anteriormente. Para el problema M_1 el conjunto de conexiones que lo modela *a priori* es de la forma:

$$L_{M1} = L^a_{M1} \cup L^i_{M1}$$

$$L^a_{M1} = \left\{\left\{u_{1j}, u_{2j}, u_o\right\}_j i = 1..3\right\} \cup \left\{\left\{u_{5,1}, u_o\right\}\right\}$$

$$L^i_{M1} = \left\{\left\{u_o\right\}\right\}$$

A priori se pueden distinguir L^a_{M1} (el conjunto de conexiones excitadoras, conexiones de peso positivo) y L^i_{M1} (el conjunto de conexiones inhibitorias, de peso negativo). Una Máquina de Boltzmann completa especificada *a priori* para el problema M_1, vendría dada por (U^{16}, L_{M1}, W_{M1}), donde

$$W_{M1}(\lambda) = \begin{cases} 4.0 & \lambda \in L^a_{M1} \\ -1.0 & \lambda \in L^i_{M1} \end{cases}$$

Para el problema M_2 la topología *a priori* y los pesos de la máquina ideal vienen dados por:

$$L_{M2} = L^a_{M2} \cup L^{i1}_{M2} \cup L^{i2}_{M2}$$

$$L^a_{M2} = \left\{\left\{u_{i1}, u_{j1}, u_o\right\}_{i,j} \in \{1..6\}\right\}$$

$$L^{i1}_{M2} = \left\{\left\{u_{i1}, u_{j1}, u_{k1}, u_o\right\}_{i,j,k} \in \{1..6\}\right\}$$

$$L^{i2}_{M2} = \left\{\left\{u_o\right\}\right\}$$

$$W_{M2}(\lambda) = \begin{cases} 1.0 & \lambda \in L^a_{M2} \\ -3.0 & \lambda \in L^{i1}_{M2} \\ -0.5 & \lambda \in L^{i2}_{M2} \end{cases}$$

Por último, para M_3 la topología y pesos de una máquina *a priori* serían:

$$L_{M3} = L^{a1}_{M3} \cup L^{a2}_{M3} \cup L^{i}_{M3}$$

$$L^{a1}_{M3} = \{\{u_{4,1}, u_{5,3}, u_o\}\}$$

$$L^{a2}_{M3} = \{\{u_o\}\}$$

$$L^{i}_{M3} = \{\{u_{5,4}, u_o\}, \{u_{2,3}, u_o\}\}$$

$$W_{M3}(\lambda) = \begin{cases} 4.0 & \lambda \in L^{a1}_{M3} \\ 1.0 & \lambda \in L^{a2}_{M3} \\ -2.0 & \lambda \in L^{i}_{M3} \end{cases}$$

Nótese que para M_1 la topología *a priori* es de orden 3, para M_2 es orden 4, y para M_3 es de orden 3. Esto indica que no podemos esperar buenos resultados en el aprendizaje aplicado a máquinas con topologías de orden inferior. Si bien siempre queda la posibilidad de que el algoritmo de aprendizaje encuentre una máquina mejor.

En los experimentos que siguen, denominamos máquina densamente conectada de orden **r** a aquellas en las que están definidas todas las conexiones *interesantes* hasta las de orden **r**. Dado que el objetivo es la clasificación de unos vectores de características, las únicas conexiones que consideraremos *interesantes* son las que unen las unidades input con las output, no consideraremos conexiones que no tengan como uno de sus extremos una unidad output. Tampoco existirán conexiones que tengan como extremos unidades que representan valores alternativos de la misma variable. Más formalmente el conjunto de conexiones de una máquina densamente conectada de orden **r** sobre el conjunto de unidades U^{16} puede expresarse como:

$$L^r = \left\{\lambda \subset U \middle| (|\lambda| \le r) \wedge (u_o \in \lambda) \wedge \left(u_{ij} \in \lambda \Rightarrow u_{ik} \notin \lambda\right)\right\}$$

Como ya se indicó en la introducción, la convexidad de la distancia de Kullback-Leibler nos permite, en todos los casos, iniciar los pesos a **cero**. Esta elección simplifica la interpretación de los resultados (no dependen del estado inicial). La regla de adaptación de los pesos es una burda aproximación del gradiente:

$$\Delta\omega_\lambda = \begin{cases} 1 & (p'_\lambda - p_\lambda) > 0 \\ -1 & (p'_\lambda - p_\lambda) < 0 \end{cases}$$

Donde p'_λ es la estimación, a partir de los datos de entrenamiento, de la probabilidad de activación de la conexión λ en la fase fijada y p_λ la de la fase libre. En la fase libre sólo la unidad output (y las unidades ocultas en su caso) evolucionan libremente. Para la estimación de las probabilidades de activación en la fase libre se realizó un enfriamiento estadístico desde temperatura 5 hasta 1.5. A temperatura 1.5 se realizó una cadena de intentos para tomar las estadísticas.

Cuando la topología es capaz de modelar la distribución de los datos, el error $\varepsilon^2 = \sum_{\lambda} \left(p'_\lambda - p_\lambda\right)^2$ converge generalmente hacia cero, y tiene sentido utilizar el error como referencia para la aplicación de algún algoritmo de poda.

En los experimentos de esta sección hemos utilizado dos mecanismos de simplificación topológica. El primero consiste en aplicar un criterio aproximado de significación estadística a las conexiones una vez que el aprendizaje está sufucientemente avanzado. El criterio utilizado consiste en eliminar aquellas conexiones cuyo peso promedio, estimado sobre los valores que toma a lo largo del aprendizaje, cae por debajo de su varianza, estimada de la misma forma. Este criterio asume la normalidad de la distribución de los pesos, eliminando aquellos que, con alta probabilidad, tienen media cero [5, 9]. En los experimentos se intentó determinar el momento apropiado para realizar esta poda en función del error.

El segundo método de simplificación topológica consistió en aplicar un término de decaimiento a la regla de adaptación de los pesos [12, 13]. De esta forma la regla aplicada es : $\Delta\omega_\lambda^d = \Delta\omega_\lambda - \theta\omega_\lambda$. Además se realiza una poda al terminar el aprendizaje, en la que se eliminan las conexiones con peso de magnitud menor que 1.

Las máquinas con unidades ocultas son de orden 2, completamente conectadas entre capas. Por consistencia, se han considerado 3,2 y 4 unidades ocultadas para los problemas M_1, M_2 y M_3 respectivamente. Estas topologías son idénticas a las utilizadas en [21] con la red entrenada con el algoritmo de contrapropagación. En todos los casos (excepto en el aprendizaje con decaimiento) las máquinas de alto orden obtienen un % de aciertos sobre el conjunto de entrenamiento es el 100%, por lo tanto la MB se ajusta siempre con precisión arbitraria al conjunto de entrenamiento, lo que no ocurre con las máquinas de orden dos con unidades ocultas.

	M1		M 2		M3	
	%aciertos	#conex	%aciertos	#conex	%aciertos	#conex
Mejor resultado [21]	100	58	100	41	100	-
Máquina *a priori*	100	5	99,8	36	100	4
Topología *a priori*	100	5	96,75	36	97,22	4
L^3	100	106	60	106	92.43	106
poda $\varepsilon^2 < 0.1$	100	17	51,8	0	52.77	1
0.01	100	22	-		96.99	33
0.001	100	23	-		92.59	34
decaimiento θ=0.1					95,37	29
L^4	-		72.68	380	93,75	380
poda $\varepsilon^2 < 0.1$	-		67.12	76	51,8	1
0.01	-		72.45	145	94.9	128
0.001	-		72.68	380	94.9	135
decaimiento θ=0.1			58,56	380	93,98	130
L^5	-		71.99	821	93.25	821
poda $\varepsilon^2 < 0.1$	-		64.81	255	47.22	1
0.01	-		68.05	313	86.80	89
0.001	-		69.67	330	87.50	99
decaimiento θ=0.1			68,98	326	94,67	93
L^6	-		70,86	1172	96,75	1172
poda $\varepsilon^2 < 0.1$	-		65,7	264	91,2	58
0.01	-		-		53,24	1
0.001	-		-		-	
decaimiento θ=0.1	-		70,3	804	93,28	774
L^7	-		71,9	1280	96,29	1280
unidades ocultas	91,3	55	67,12	36	96,75	75

Tabla 1. Resultados con unidades binarias para la topología *a priori* y topologías densamente conectadas, con aplicación de algoritmos de simplificación

La tabla 1 resume los resultados de los experimentos realizados. Por filas, la tabla recoge: el mejor resultado reportado en [21] (#con es el numero de conexiones cuando el vencedor fue el algoritmo de retropropagación), los resultados con la máquina *a priori* (sin aprendizaje de pesos), con la topología *a priori* (sólo se ajustan los pesos de las conexiones establecidas *a priori*), y con las topología densamente conectadas de orden creciente. Sobre las topologías densamente conectadas se realizan el aprendizaje puro, y las simplificaciones topológicas aplicando el criterio de significación estadística (poda) y el decaimiento de pesos. La activación del criterio estadístico de poda se realizó al alcanzar distintos niveles de error. En el problema M1 la experimentación se detuvo en las máquinas de orden 3, mientras que para los otros dos problemas se extendió hasta las máquinas completamente conectadas (orden 7).

Puede observarse que las topologías *a priori* producen los mejores resultados con el mínimo de conexiones. El problema M_2 resulta especialmente difícil, en gran parte debido a que la distancia mínima entre patrones dentro y fuera de la clase a detectar es 1. Como ya comentábamos en la introducción, la dificultad de este problema nos indujo

a la introducción de unidades no binarias en la formulación de las Máquinas de Boltzmann que se discute en la próxima sección.

Los esquemas de poda y decaimiento no producen, en general, mejoras en el número de aciertos sobre el conjunto test, pero si disminuciones notables del número de conexiones. El problema substancial de la poda es el descubrimiento de la topología ideal, que es la más simple y la que proporciona los mejores resultados. En este sentido los algoritmos de poda probados distan mucho de cumplir este objetivo.

3 Aprendizaje en Máquinas de Boltzmann con unidades no binarias

La notación para las MB con unidades de estados discretos es una extensión de la del caso binario, añadiendo la especificación de los rangos de valores que pueden tomar los estados de las unidades. Una máquina de Boltzmann de unidades discretas viene descrita por la cuádrupla (U,R,L,W) en la que U, L, W mantienen su significado y $R = \{R_i \subset Z\}$ donde R_i es el espacio de estados de la unidad u_i. Mantenemos la interpretación multiplicativa de las conexiones, por tanto, la función de consenso mantiene la forma $C(k) = \sum_{\lambda} \omega_\lambda \prod_{u \in \lambda} k(u)$ con la precisión de que $k(u_i) \in R_i$.

Siguiendo la linea de análisis de Aarts para las máquinas binarias de orden 2, puede demostrarse que para que una distribución de probabilidad **q'** sea realizable por la máquina de Boltzmann de alto orden con unidades discretas, es condición necesaria que los rangos de las unidades contengan el valor 0, esto es: $\forall R_i \in R, 0 \in R_i$.

Además, la regla de aprendizaje de los pesos puede calcularse (ver Apéndice 1) como es habitual como un descenso por el gradiente de la distancia de Kullback-Leibler D(**q'**/**q**) entre la distribución deseada **q'** y **q**, la que ofrece la máquina (a temperatura 1). El gradiente resulta ser de la forma:

$$\frac{\partial D(\mathbf{q}'/\mathbf{q})}{\partial \omega_\lambda} = -\frac{1}{c}(a'_\lambda - a_\lambda)$$

donde $a_\lambda = \sum_k q_k \prod_{u \in \lambda} k(u)$ es el nivel de activación promedio de la conexión λ bajo una distribución $\mathbf{q}_k$ de la configuraciones, y c es el parámetro de temperatura. Condiciones de convergencia similares a las dadas en [3] también se obtienen para este caso (ver Apéndice 1), considerando las segundas derivadas de la distancia D respecto de los pesos, para el caso de que los rangos de valores de las unidades sean positivos: $R_i \subset N \cup \{0\}$. También en este caso es posible garantizar la convexidad de la distancia de Kullback-Leibler.

Consideramos un problema abierto la caracterización (de forma análoga a lo realizado en [3]) de las distribuciones de probabilidad que la máquina de Boltzmann de alto orden

con unidades discretas es capaz de aprender, y la relación con su topología. También merece estudio más detallado, por sus implicaciones prácticas, la convexidad de la distancia de Kullback-Leibler en el caso de que los rangos de valores de las unidades incluyan valores negativos.

4 Resultados de las Máquinas de Boltzmann no binarias aplicadas los problemas de MONK's

El conjunto de unidades utilizado para representar los datos es $U^7=\{u_i\ i=1..6,\ u_o\}$, los rangos de las unidades son de la forma: $R_1=R_2=R_4=\{0..2\}$, $R_3=R_6=R_o=\{0..1\}$ y $R_5=\{0..3\}$. Las topologías *a priori* vienen definidas por las relaciones de pesos que proporcionan máximo consenso para las configuraciones deseadas. Encontrar estas topologías no parece evidente para los problemas M_1 y M_3. Sin embargo si es posible encontrarla para M_2. Esta topología y pesos ideales son de la forma

$$L_{M2} = L^{a}_{M2} \cup L^{i1}_{M2} \cup L^{i2}_{M2}$$

$$L^{a}_{M2} = \left\{\left\{u_i, u_j, u_k, u_m, u_o\right\}, j,k,m \in \{1..6\}\right\}$$

$$L^{i1}_{M2} = \left\{\left\{u_i, u_j, u_k, u_m, u_n, u_o\right\}, j,k,m,n \in \{1..6\}\right\}$$

$$L^{i2}_{M2} = \left\{\left\{u_o\right\}\right\}$$

$$W_{M2}(\lambda) = \begin{cases} 2.0 & \lambda \in L^{a}_{M2} \\ -30.0 & \lambda \in L^{i1}_{M2} \\ -1.0 & \lambda \in L^{i2}_{M2} \end{cases}$$

También en este caso realizamos experimentos con máquinas densamente conectadas y algoritmos de simplificación topológica. Las máquinas densamente conectadas de orden **r** , definidas sobre el conjunto de unidades U^7, tienen las siguientes conexiones

$$L^{r} = \left\{\lambda \subset U^7 \middle| (|\lambda| \le r) \wedge (u_o \in \lambda)\right\}$$

	M1		M 2		M3	
	%aciertos	#con	%aciertos	#con	%aciertos	#con
Mejor resultado [21]	100	58	100	41	100	-
A priori HOBM	-		100	22	-	
Topología *a priori*	-		95	22	-	
L^3	77	22	64	22	88	22
L^4	80	42	75	42	89	42
L^5	81	57	84	57	85	57
poda $\varepsilon^2 < 0.1$	81	24	67	9	-	
L^6	81	63	91,9	63	91,8	63
poda $\varepsilon^2 < 0.1$	81	63	54	13	79	25
0.01	-		88	24	82	
0.001	81	63	-		86	
decaimiento θ=0.1	50	1	50	15	60	

•

- Tabla 2. Resultados con máquinas de alto orden con unidades discretas

La tabla 2 muestra los resultados de la aplicación de las máquinas de Boltzmann de alto orden con unidades discretas a los problemas de MONK's. Al igual que en la sección 2, los pesos iniciales son siempre **cero**. La regla de adaptación de los pesos tiene ahora la forma:

$$\Delta\omega_\lambda = \begin{cases} 1 & (a'_\lambda - a_\lambda) > 0 \\ -1 & (a'_\lambda - a_\lambda) < 0 \end{cases}$$

Los algoritmos de poda y decaimiento son idénticos a los ya presentados en la sección 2. Las cantidades a'_λ y a_λ son estimaciones de las activaciones promedio de las conexiones en las fases fijada y libre. El proceso de obtención de estas activaciones promedio se realiza utilizando un enfriamiento estadístico similar al de la sección 2.

Se observa que el problema M_2, para el que las máquinas binarias daban los peores resultados, se resuelve satisfactoriamente. Los algoritmos de poda y decaimiento no ofrecen buenos resultados en general. También se observa que, a pesar de que no parece existir una topología *a priori* para los problemas M_1 y M_3, el algoritmo de aprendizaje es capaz de obtener aproximaciones razonablemente buenas.

5 Aplicación de las Máquinas de Boltzmann de alto orden al problema de reconocimiento de señales sonar

Los datos que definen este problema son los utilizados por Gorman y Sejnowski [10] en su estudio sobre la clasificación de señales sonar utilizando redes entrenadas con el algoritmo de retropropagación y han sido obtenidos vía ftp de la base de datos pública de la CMU, (ftp.cs.cmu.edu, directorio /afs/cs/project/connect/bench). La tarea a

realizar consiste en la discriminación entre las señales sonar reflejadas por un cilindro de piedra y otro de metal. Los datos se dividen en un conjunto de entrenamiento y un conjunto test de 104 patrones cada uno. Cada patrón consta de 60 inputs continuos (con valores entre 0 y 1) y un output binario que indica la clase a la que pertenece el patrón input.

$$\mathbf{x} = (x_1, \ldots, x_{60}, x_o) \in [0,1]^{60} \times \{\text{metal}, \text{roca}\}$$

En [10] se recogen dos clases de experimentos dependiendo de si el ángulo de incidencia se toma en cuenta o no para definir los conjuntos de entrenamiento y test. Los datos obtenidos de la CMU están distribuidos en los conjuntos de entrenamiento y test de forma que incluyen el mismo porcentaje patrones para los distintos ángulos de incidencia. Esto corresponde al segundo conjunto de experimentos relatados en [10]. Los resultados que proporcionan los autores se obtienen promediando 10 aplicaciones del algoritmo de aprendizaje a partir de pesos aleatorios. En la regla de actualización de los pesos que aplican, se utiliza una velocidad de aprendizaje de 2.0, momento 0.0 y los errores inferiores a 0.2 se tratan como 0.0. En resumen los resultados se recogen en la tabla 3:

Unidades ocultas	%aciertos entrenamiento	Desv. Std	%aciertos test	Desv. Std..
0	79.3	3.4	73.1	4.8
2	96.2	2.2	85.7	6.3
3	98.1	1.5	87.6	3.0
6	99.4	0.9	89.3	2.4
12	99.8	0.6	90.4	1.8
24	100.0	0.0	89.2	1.4

Tabla 3. Resultados de Gorman y Sejnowski para los datos distribuidos uniformemente según el ángulo de incidencia.

En primera aproximación intentamos aplicar a este problema las Máquinas de Boltzmann de alto orden con unidades discretas. El conjunto de unidades de la máquina es $U=\{u_i \; i=1..60, \; u_o\}$, donde los rangos de valores de las unidades son $R_1=..=R_{60}=\{0..10000\}$ y $R_o=\{0..1\}$. Las unidades input toman el valor del input correspondiente multiplicado por 10^4: $k(u_i)=x_i*10^4$ (La precisión de los datos es de 4 dígitos). La unidad output toma valor 0 si la clase es metal y 1 si es roca. Las topologías experimentadas con esta aproximación fueron de orden 2 y orden 3 densamente conectadas (obviamente no conocemos ninguna topología *a priori*).

$$L^2 = \left\{\lambda \subset U \middle| (|\lambda| \leq 2) \wedge (u_o \in \lambda)\right\}$$

$$L^3 = \left\{\lambda \subset U \middle| (|\lambda| \leq 3) \wedge (u_o \in \lambda)\right\}$$

Los resultados que se recogen en la tabla 4 muestran la dificultad de definir los principales parámetros involucrados en el aprendizaje: la temperatura inicial y final del enfriamiento estadístico realizado en la fase libre, y la velocidad de aprendizaje. Los experimentos parten, como siempre, de pesos iniciales nulos, y la adaptación de los

pesos es función de los niveles de activación promedio (α es la velocidad de aprendizaje):

$$\Delta\omega_\lambda = \begin{cases} \alpha & (a'_\lambda - a_\lambda) > 0 \\ -\alpha & (a'_\lambda - a_\lambda) < 0 \end{cases}$$

La tabla 4 muestra los resultados de los experimentos realizados con las máquinas con estados discretos sobre los datos del sonar. Se aprecia en ellos las fuertes variaciones en magnitud de los parámetros del aprendizaje, y la sensibilidad del aprendizaje a sus valores. Los experimentos, como puede apreciarse, son incompletos y orientativos, nos sirven de motivación empírica para la introducción de las máquinas de Boltzmann con unidades de estados continuos. Nótese que en esta tabla mostramos también el número de ciclos de aprendizaje y los resultados obtenidos tanto sobre el conjunto de entrenamiento como sobre el conjunto de test. Para mostrar la ocurrencia de sobreajuste mostramos especificamente como para la topología L^2 con α=0.01 el pico de rendimiento sobre el conjunto test (79%) se obtiene a los 4200 ciclos descendiendo luego a los 5000 ciclos hasta el 71%.

Topología	temp inicial	temp. final	α	ciclos	%aciertos train	%aciertos test
L^2	10^5	$3*10^3$	1.0	3000	47	59
			0.01	4200	86	79
				5000	86	71
	10^6	$3*10^5$	1.0	3100	85	76
			0.01	5000	90	75
L^3	10^6	$3*10^5$	0.01	830	52	42
	10^9	$3*10^8$	1.0	1160	54	46
			0.001	550	83	78

Tabla 4. Resultados con Máquinas de Boltzmann de alto orden con unidades de estados discretos (unidades no binarias) para el reconocimiento de señales sonar.

5.1 Máquinas de Boltzmann con unidades continuas para el reconocimiento de señales Sonar

La introducción de unidades continuas únicamente varia la definición del espacio de estados (configuraciones) en el que está definida la máquina de Boltzmann. Para el problema de reconocimiento de señales sonar, los rangos de valores para las unidades input son ahora: R_1=..=R_{60}=[0..1]. Los estados de la unidades input son los de los componentes input del patrón $k(u_i)=x_i$. La unidad output toma valores 0 ó 1 como en el caso discreto. En los experimentos se utilizaron dos tipos de topologías generales: las densamente conectadas

$$L^r = \left\{\lambda \subset U \middle| (|\lambda| \le r) \wedge (u_o \in \lambda)\right\}$$

y las conectadas en linea, esto es, topologías densamente conectadas con la restricción de que las unidades input conectadas son contiguas:

$$L_{|}^{r} = \left\{ \lambda \subset U \middle| \left(|\lambda| \le r \right) \wedge \left(u_o \in \lambda \right) \wedge \left(r > 2 \Rightarrow \left(u_i \in \lambda \Rightarrow \left(u_{i-1} \in \lambda \vee u_{i+1} \in \lambda \right) \right) \right) \right\}$$

En los experimentos recogidos en la tabla 5, como siempre, los pesos iniciales son cero y la velocidad de aprendizaje es 1.0. Para las topología en linea se realiza la poda basada en la relación entre la media y la desviación típica del peso a partir del ciclo 50. En la mayor parte de los experimentos la toma de las estadísticas en la fase libre se han realizado a una temperatura (sin realizar enfriamiento desde una temperatura superior), por lo que la temperatura final (que no se especifica) coincide con la inicial. Esta simplificación reduce los tiempos de CPU considerablemente, y no afecta a los resultados del aprendizaje. Los experimentos recorren varias topologías de alto orden y se incluyen experimentos con máquinas de Boltzmann convencionales con 2, 5 y 6 unidades ocultas por consistencia experimental. Computacionalmente, las topologías con unidades ocultas son mucho más costosas, causa por la cual no nos hemos detenido a examinarlas "in extenso".

Topología	t. inicial	t.final	ciclos	%train	%test	#conex
L^2	500	150	340	80	81	61
			5000	91	78	
L^3	3000	-	5000	97	86	1084
L^4	30000		550	86	77	31084
$L_{|}^4$	200	-	5000	93	82	82
$L_{|}^5$	250	-	5000	97	87	119
$L_{|}^6$	350	-	5000	100	86	172
$L_{|}^7$	400	-	5000	99	87	209
2 ocultas	75	-	210	53	42	126
5 ocultas	150	-	5000	86	78	321
6 ocultas	180	-	4450	87	81	388

Tabla 5. Resultados sobre el reconocimiento de señales sonar con máquinas de Boltzmann con unidades continuas.

Se observa en la tabla 5 que el aprendizaje reduce su sensibilidad a los parámetros de temperatura y velocidad de aprendizaje. La máquina de orden 2 se comporta de forma comparable a la red sin unidades ocultas de Gorman y Sejnowski. La topología densamente conectada de orden 3 mejora los resultados pero su costo comienza a ser elevado. La topología de orden 4 es tan costosa que tuvimos que detener prematuramente el proceso de aprendizaje. Las topologías en linea presentan un comportamiento excelente, son en cierta medida las topologías ideales o *a priori* para el problema, proporcionan el mejor ajuste y generalización con el mínimo costo computacional. Las correlaciones entre los inputs más cercanos en el tiempo y la situación de los picos de la señal parecen ser las características mas influyentes en la discriminación.

6 Aplicación a un problema de reconocimiento de vocales

El problema consiste en el reconocimiento de las vocales inglesas. Los datos de este problema han sido tomados también de la base de datos pública de la CMU (ftp.cs.cmu.edu). Los datos fueron utilizados por T. Robinson (originalmente recogidos por D. Deterding [8]) en la realización de su tesis doctoral [18,19], la cual incluia la comparación del rendimiento de distintas arquitecturas sobre este conjunto de datos. Los mejores resultados que informa son producidos por la clasificación por el vecino más cercano basados en la distancia euclídea: clasifica correctamente un **56**% del conjunto de test. Cada uno de los patrones vienen dados por 10 inputs reales y un output que identifica la clase (vocal)

$$\mathbf{x} = (x_1, \ldots, x_{10}, x_o) \in R^{10} \times \text{Vowels}$$

$$\text{Vowels} = \{\text{hid}, \text{hId}, \text{hEd}, \text{dAd}, \text{hYd}, \text{had}, \text{hOd}, \text{hod}, \text{hUd}, \text{hud}, \text{hed}\}$$

Los datos recogen muestras de las once vocales del inglés pronunciadas por quince hablantes distintos. Cuatro hombres y cuatro mujeres produjeron los datos en el conjunto de entrenamiento. Cuatro hombres y tres mujeres los de test. El detalle de la obtención de los datos puede encontrarse en [18,19] y en la descripción del problema que acompaña a los datos. El conjunto de entrenamiento está dado por 528 patrones y el de entrenamiento por 462.

Las características que hacen interesante este problema, en relación a los anteriores, son las siguientes:

Es un problema de clasificación multicategórica

Los datos input no están normalizados en el intervalo [0,1], en general x_i [-5,5].

Los estados de las unidades input incluyen valores negativos, por tanto no está garantizada la convexidad de la distancia de Kullback-Leibler. Es importante comprobar que incluso en esta situación el algoritmo de aprendizaje se comporta bien

El conjunto de unidades que consideramos para este problema es $U=\{u_i\ i=1..10,\ u_{oj}\ j=1..11\}$, las unidades input u_i tienen rangos $R_i=[-5,5]$, las unidades output u_{oj} tienen rango $R_{oj}=\{0,1\}$. Las unidades input toman el valor del componente input sin normalizar $k(u_i)=x_i$. Las unidades output toman valor 0 ó 1 $k(u_{oj})=1$ si x_o toma el j-esimo valor en su rango. Las topologías experimentadas han sido las densamente conectadas en las que siempre hay una única unidad ouput en la conexión:

$$L^r = \left\{ \lambda \subset U \middle| \left(|\lambda| \leq r \right) \wedge \left(u_{oj} \in \lambda \right) \wedge \left(\forall k \neq j \left(u_{ok} \notin \lambda \right) \right) \right\}$$

y las densamente conectadas en linea.

$$L_l^r = \left\{ \lambda \in L^r \middle| \left(r > 2 \Rightarrow \left(u_i \in \lambda \Rightarrow \left(u_{i-1} \in \lambda \vee u_{i+1} \in \lambda \right) \right) \right) \right\}$$

La tabla 6 recoge los resultados obtenidos con el esquema convencional de aprendizaje: incrementos fijos de los pesos y toma de estadísticas en la fase libre basada en la simulación de la máquina a una temperatura. Debido a la dificultad de determinar la temperatura apropiada y a la ausencia de unidades ocultas, decidimos realizar la estimación de las activaciones promedio de las conexiones en la fase libre de forma aproximada. La respuesta en las unidades ouput se calcula buscando directamente la unidad con input neto máximo. Esta aproximación elimina el ruido introducido por la búsqueda aleatoria. Algunos autores [1, 2, 3] afirman que este ruido puede ser beneficioso para ayudar a algoritmo de aprendizaje a escapar de mínimos locales. Nuestra experiencia en este problema concreto contradice esas afirmaciones. Recuerdese que la ausencia de unidades ocultas hace que sea convexa (en el caso binario) la distancia que minimiza el aprendizaje. En este problema la convexidad es una asunción dudosa, que, sin embargo, se ve confirmada por los resultados. Asumida la convexidad del espacio de búsqueda, no existen mínimos locales de los que el ruido deba sacar al algoritmo de aprendizaje. La tabla 7 recoge los resultados del aprendizaje sin relajación estocástica, con incremento fijo de pesos.

Topología	t. inicial	t.final	ciclos	%train	%test	#conex
L^2	300	-	1630	15	7	176
L^3	120	-	230	33	19	671
L^4	500	-	380	52	25	1991
L^5	400	-	210	45	23	4301
L_l^3	600	-	1040	35	17	275
L_l^4	900	-	830	34	14	363
L_l^5	1200	-	330	24	14	440
L_l^6	1200	-	260	22	14	

Tabla 6. Resultados sobre el reconocimiento de vocales con máquinas de Boltzmann con unidades continuas, adaptación de los pesos ±1 y estimación de los niveles de activación usando relajación estocástica a temperatura fija.

Topología	ciclos	%train	%test
L^2	180	28	19
L^3	100	55	30
	340	65	31
L^4	140	72	37
	610	89	36
L^5	30	72	34
	570	90	39
	890	93	33
$L_{\|}^3$	50	43	29
	610	55	29
$L_{\|}^4$	30	37	30
	640	57	26
$L_{\|}^5$	100	55	31
	1000	62	25
$L_{\|}^6$	100	49	20
	1000	67	28
$L_{\|}^7$			
$L_{\|}^8$	100	55	24
	1000	66	25

Tabla 7. Resultados sobre el reconocimiento de vocales con máquinas de Boltzmann con unidades continuas, incremento de pesos ±1 y estimación de las activaciones promedio sin relajación estocástica

Debido a los pobres resultados obtenidos, en relación a los de referencia, probamos dos reglas más depuradas de actualización de los pesos. La primera utilizando directamente el gradiente:

$$\Delta\omega_\lambda = \left(a'_\lambda - a_\lambda\right)$$

La segunda utilizando un término de momento:

$$\Delta_t\omega_\lambda = \left(a'_\lambda - a_\lambda\right) + \mu\Delta_{t-1}\omega_\lambda$$

Las tablas 8 y 9 reflejan los resultados con ambas reglas. Utilizando el término de momento conseguimos obtener resultados similares a los de referencia.

Topología	ciclos	%train	%test
L^2	30	32	24
	200	38	30
	400	48	37
L^3	30	52	31
	200	69	43
	230	78	52
	400	88	50
	1000	99	50
L^4	30	67	41
	100	88	46
	200	99,8	45
L^5	30	89	46
	60	99	45
	130	100	45
L^6	30	90	41
	80	100	41

Tabla 8. Resultados sobre el reconocimiento de vocales con máquinas de Boltzmann con unidades continuas, adaptación de pesos $\left(a'_\lambda - a_\lambda\right)$ y estimación de las activaciones promedio sin relajación estocástica

Se observa en los resultados la mejora debida al refinamiento del algoritmo de aprendizaje. Los ciclos en los que se dan resultados se han escogido para hacer patente el fenómeno de sobreajuste. Este fenómeno se produce con mayor claridad en las máquinas de orden más elevado. En este caso, el nivel de aciertos sobre el conjunto test alcanza un pico y desciende. Si el orden no es demasiado elevado para el problema los aciertos sobre el conjunto test se mantienen oscilando cerca del valor pico. En todo caso, la máquina tiende a ajustar el conjunto de entrenamiento al 100%. El mejor resultado se obtiene en la tabla 9 para la topología en linea de orden 3, donde el valor pico sobre el conjunto test supera a los de referencia (58% frente a 56%).

Atendiendo al tipo de topología, se observa también que las topología en linea funcionan en general de forma muy eficiente a pesar de su simplicidad. De hecho, el mejor resultado lo proporciona una de ellas. No creemos que estos resultados sean extrapolables, en el sentido de que las topologias en linea sean universalmente apropiadas. Sin embargo resultan chocantes, puesto que nada hace preveer que el problema pueda ajustarse por estas topologías.

La eliminación de la relajación estocástica y la aplicación de reglas de adaptación de los pesos producen no sólo mejor calidad de los resultados, sino una aceleración substancial en este caso. Esta aceleración puede apreciarse comparando los ciclos en los que se dan resultados en las diferentes tablas. No consideramos la medición de los tiempos absolutos de CPU, puesto que estos son tremendamente dependientes de la máquina e implementación concreta.

Topología	ciclos	%train	%test
L^2	30	39	33
	100	62	43
	150	59	39
L^3	30	69	47
	80	98	54
	200	99	53
L^4	30	89	44
	40	96	46
	100	99	45
	390	99	41
L^5	30	94	42
	60	100	45
L^6	30	97	40
	50	100	45
$L^3_{\mid}$	30	38	24
	120	87	**58**
	250	89	54
$L^4_{\mid}$	30	43	27
	120	87	57
	250	95	54
$L^5_{\mid}$	30	46	27
	120	84	51
	250	91	51
$L^6_{\mid}$	30	53	39
	120	83	52
	250	97	55
$L^7_{\mid}$	30	45	27
	120	78	49
	250	90	53
$L^8_{\mid}$	30	57	27
	120	81	50
	250	95	53
$L^9_{\mid}$	30	56	30
	120	92	51
	150	96	52

Tabla 9. Resultados sobre el reconocimiento de vocales con máquinas de Boltzmann con unidades continuas, adaptación de pesos $\left(a'_\lambda - a_\lambda\right) + \mu\Delta_{t-1}\omega_\lambda$ con μ=0.9 y estimación de las activaciones promedio sin relajación estocástica

7 Conclusiones y trabajo futuro

En este artículo introducimos las Máquinas de Boltzmann de alto orden. Se han realizado experimentos que demuestran que el aprendizaje con esta arquitectura neunonal es robusto y eficiente. El algoritmo se generaliza unidades discretas con rango general y a unidades continuas. Los experimentos se han descrito siguiendo la secuencia histórica de nuestros trabajos, que nos lleva a una versión altamente eficiente y de muy buenos resultados. En general, los experimentos involucran topologías de orden creciente, algo no realizado previamente en la literatura. Nuestra conclusión es que las Máquinas de Boltzmann de alto orden son una arquitectura eficiente y de aprendizaje rápido y robusto, que poseen un potencial de aplicación práctica similar, al menos, al de otras arquitecturas neuronales. Continuamos realizando experimentos de aplicación a conjuntos de datos de dominio público, buscando aumentar nuestra confianza en su aplicabilidad práctica.

En general, se observa con mucha claridad en el proceso de aprendizaje el fenómeno del sobreajuste. Conforme progresa el aprendizaje, el ajuste sobre el conjunto de entrenamiento mejora hasta el límite impuesto por la propia topología y los resultados sobre el conjunto de test se estabilizan por debajo del mejor resultado obtenido durante el proceso de aprendizaje. Las Máquinas de Boltzmann de alto orden con topologías densamente conectadas, caen muy fácilmente en sobreajuste. El problema de la simplificación o determinación topológica parece especialmente crítico si se intenta aplicar estas arquitecturas a problemas reales. En este trabajo se han intentado dos aproximaciones a este problema: la poda utilizando criterios estadísticos y el decaimiento de pesos. Ninguna de las dos técnicas ha proporcionado resultados inmediatos que resulten definitivos. Sin embargo, creemos que el refinamiento de estos métodos puede conducir a resultados generales que hagan todavía más interesantes y útiles las máquinas de alto orden.

La realización de la toma de estadísticas obviando el enfriamiento estadístico está condicionada a la ausencia de unidades ocultas y a la clase de problemas que se han tratado en este trabajo. Estos son problemas de clasificación donde el output de la red viene dado por un conjunto de unidades binarias, y donde las respuestas deseadas de la red son ortogonales entre si. La respuesta consiste siempre en un vector de ceros y un único uno. En estas condiciones la búsqueda de la respuesta dado un input es una búsqueda de complejidad lineal de la unidad output con máximo incremento de consenso respecto de la respuesta nula (todo ceros). No es obvio que esta aproximación pueda extenderse a otros casos. Sin embargo, esta clase de problemas es lo suficientemente amplia y de interés práctico, como para ser tomada en cuenta.

Consideramos problemas abiertos la caracterización de las distribuciones de probabilidad que las Máquinas de Boltzmann de estados discretos (no binarios) y continuos, así como la justificación rigurosa del algoritmo de aprendizaje para el caso de las unidades continuas. También merece un estudio más detallado, por sus implicaciones prácticas, la convexidad de la distancia de Kullback-Leibler, entre la distribución a aprender y la de la máquina, en el caso más general, en el que las unidades pueden tomar estados negativos. Nuestra experiencia con el problema de las vocales parece indicar que esta convexidad existe en condiciones muy generales.

Referencias

[1] Aarts E.H.L., J.H.M. Korst "Simulated Annealing and Boltzmann Machines: a stochastic approach to combinatorial optimization and neural computing" John Wiley & Sons (1989)

[2] Ackley D.H., G.E. Hinton, T.J. Sejnowski "A learning algorithm for Boltzmann Machines" Cogn. Sci. 9 (1985) pp.147-169

[3] Albizuri F.X. , A. D´Anjou, M. Graña, F.J. Torrealdea, M.C. Hernandez "The High Order Boltzmann Machine: learned distribution and topology" IEEE Trans. Neural Networks en prensa

[4] Albizuri F.X. Tesis Doctoral en preparación, Dept. CCIA Univ. Pais Vasco

[5] Cottrell M et alt. (1993) "Time series and neural networks: a statistical method for weight rlimination" ESSAN'93 pp.157-164

[6] D´Anjou A., M. Graña, F.J. Torrealdea, M.C. Hernandez "Máquinas de Boltzmann para la resolución del problema de la satisfiabilidad en el cálculo proposicional" Revista Española de Informática y Automática 24 (1992) pp.40-49

[7] D´Anjou A., M. Graña, F.J. Torrealdea, M.C. Hernandez "Solving satisfiability via Boltzmann Machines" IEEE Trans. on Patt. An. and Mach. Int. Mayo 93

[8] Deterding D. H. (1989) "Speaker Normalisation for Automatic Speech Recognition", PhD Thesis, University of Cambridge,

[9] Fambon O., C Jutten (1994) "A comparison of two weight pruning methods" ESSAN'94 pp.147-152

[10]. Gorman, R. P., and Sejnowski, T. J. (1988). "Analysis of Hidden Units in a Layered Network Trained to Classify Sonar Targets" in Neural Networks, Vol. 1, pp. 75-89.

[11] Graña M. , V. Lavin, A. D'Anjou, F.X. Albizuri, J.A. Lozano "High-order Boltzmann Machines applied to the MONK's problems" ESSAN'94, DFacto press, Bruselas, Bélgica, pp117-122

[12] Hertz J., A.Krogh, R.G.Palmer "Introduction to the theory of Neural Computation". Addison Wesley, 1991.

[13] Hinton G.E. , Lectures at the Neural Network Summer School, Wolfson College, Cambridge, Sept. 1993

[14] Kohonen T. "Self-Organization and Associative Memory". Springer-Verlag 1989

[15] Kosko B. "Neural Networks and Fuzzy Systems". Prentice Hall, 1992.

[16] Perantonis S.J. , P.J.G. Lisboa "Translation, rotation and scale invariant pattern recognition by high-order neural networks and moment classifiers" IEEE Trans. Neural Net. 3(2) pp.241-251

[17] Pinkas G. "Energy Minimization and Satisfiability of Propositional Logic" en Touretzky, Elman, Sejnowski , Hinton (eds) 1990. Connectionist Models Summer School

[18] Robinson A. J. (1989) "Dynamic Error Propagation Networks".PhD Thesis Cambridge University Engineering Department,

[19] Robinson A. J. , F. Fallside (1988), "A Dynamic Connectionist Model for Phoneme Recognition" Proceedings of nEuro'88, Paris, June,

[20] Sejnowski TJ. "Higher order Boltzmann Machines" in Denker (ed) Neural Networks for computing AIP conf. Proc. 151, Snowbird UT (1986) pp.398-403
[21] Thrun S.B. et alters "The MONK´s problems: A performance comparison of different learning algorithms" Informe CMU-CS-91-197 Carnegie Melon Univ.
[22] Zwietering P.J. , Aarts E.H.L. "The Convergence of Parallel Boltzmann Machines" en Eckmiller, Hartmann, Hauske (eds) Parallel Processing in Neural Systems and Computers North-Holland 1990 pp277-280
[23] Zwietering P.J. , Aarts E.H.L "Parallel Boltzmann Machines: a mathematical model" Journal of Parallel and Dist. Computers 13 pp.65-75

Apendice 1: Derivación del algoritmo de aprendizaje de la Máquina de Boltzmann con unidades de estados discretos (no necesariamente binarios)

La definición formal de la Máquina de Boltzmann de alto orden con estados discretos

HOBMD = (U,R,L,W), $U = \{u_1,\ldots,u_n\}$ las unidades de la máquina

$k:U \to R$ estado de las unidades

$R = \underset{i=1}{\overset{n}{\times}} R_i \qquad R_i \subset Z$ t.q. $k(u_i)$ R_i espacio de estados

$L \subset 2^U$ las conexiones entre unidades (λ)

$W:L\varnothing R$ los pesos de las conexiones (ω_λ)

El número de configuraciones globales de la máquina será: $|R| = \prod_{i=1}^{n} |R_i|$

El nivel de activación de una conexión lo definimos como $a_\lambda(k) = \prod_{i\in\lambda} k(u_i)$

La función de consenso se define $C(k) = \sum_{\lambda\in L} \omega_\lambda a_\lambda(k)$

La distribución de equilibrio de la Maquina de Boltzmann a temperatura c es precisamente la distribución de Boltzmann:

$$\mathbf{q}_k(c) = \frac{e^{\frac{C(k)}{c}}}{z}; \qquad z = \sum_{k\in R} e^{\frac{C(k)}{c}}$$

El comportamiento global a temperatura c es la distribución de probabilidad conjunta:

$$\mathbf{q}(c) = \left(\mathbf{q}_1(c), \mathbf{q}_2(c), \ldots, \mathbf{q}_{|\Re|}(c)\right)$$

Para un comportamiento global, el promedio de activacion de una conexión es

$$\bar{a}_\lambda = \sum_{k\in\Re} \mathbf{q}_k(c) a_\lambda(k)$$

y la probabilidad de activación de una conexión es

$$p_\lambda = \sum \mathbf{q}_k(c)\delta(a_\lambda(k)) \qquad \delta(x) = \begin{cases} x \neq 0 \\ 0 \quad x = 0 \end{cases}$$

Como es habitual para las Máquinas de Boltzmann, el algoritmo de aprendizaje se define como la minimización, mediante descenso del gradiente, de la llamada distancia de Kullback-Leibler

$$D(\mathbf{q}/\mathbf{q}') = \sum_{k\in R} \mathbf{q}'_k(c) \ln \frac{\mathbf{q}'_k(c)}{\mathbf{q}_k(c)}$$

La derivada de dicha distancia,respecto de los pesos se calcula, en el caso de unidades de estados discretos, como sigue:

$$\frac{\partial D(\mathbf{q}/\mathbf{q}')}{\partial \omega_\lambda} = \frac{\partial}{\partial \omega_\lambda} \sum_{k\in R} \mathbf{q}'_k(c)\left(\ln \mathbf{q}'_k(c) - \ln \mathbf{q}_k(c)\right) = -\sum_{k\in R} \frac{\mathbf{q}'_k(c)}{\mathbf{q}_k(c)} \frac{\partial}{\partial \omega_\lambda} \mathbf{q}_k(c)$$

$$\frac{\partial}{\partial \omega_\lambda} \mathbf{q}_k(c) = \frac{\partial}{\partial \omega_\lambda} \frac{e^{\frac{C(k)}{c}}}{z} = \frac{\frac{1}{c} z e^{\frac{C(k)}{c}} a_\lambda(k) - \frac{1}{c} e^{\frac{C(k)}{c}} \sum_{k\in\Re} e^{\frac{C(k)}{c}} a_\lambda(k)}{z^2} =$$

$$= \frac{1}{c} \mathbf{q}_k(c) a_\lambda(k) - \frac{1}{c} \mathbf{q}_k(c) \sum_{k\in\Re} \mathbf{q}_k(c) a_\lambda(k) =$$

$$= \frac{1}{c} \mathbf{q}_k(c)\left(a_\lambda(k) - \bar{a}_\lambda\right)$$

$$\frac{\partial D(\mathbf{q}/\mathbf{q}')}{\partial \omega_\lambda} = -\sum_{k\in R} \frac{\mathbf{q}'_k(c)}{\mathbf{q}_k(c)} \frac{1}{c} \mathbf{q}_k(c)\left(a_\lambda(k) - \bar{a}_\lambda\right) = -\frac{1}{c}\left(\sum_{k\in R} \mathbf{q}'_k(c) a_\lambda(k) - \sum_{k\in R} \mathbf{q}'_k(c) \bar{a}_\lambda\right) =$$

$$= -\frac{1}{c}\left(\bar{a}'_\lambda - \bar{a}_\lambda\right)$$

Las condiciones de convergencia del descenso por el gradiente $\Delta\omega_\lambda = -\alpha \frac{\partial D(\mathbf{q}/\mathbf{q}')}{\partial \omega_\lambda}$ se determinan a partir de las segundas derivadas. En general, para Máquinas de Boltzmann sin unidades ocultas, la distancia de Kullback-Leibler es convexa y podemos deducir un valor de α que garantiza la convergencia del descenso por el gradiente. La expresión de las segundas derivadas respecto de los pesos se calcula como sigue:

$$\frac{\partial^2 D(\mathbf{q}/\mathbf{q}')}{\partial\omega_\mu\omega_\lambda} = \frac{\partial}{\partial\omega_\mu}\left[-\frac{1}{c}\left(\bar{a}'_\lambda - \bar{a}_\lambda\right)\right] = \frac{1}{c}\frac{\partial}{\partial\omega_\mu}\bar{a}_\lambda = \frac{1}{c}\frac{\partial}{\partial\omega_\mu}\sum_{k\in R}\mathbf{q}_k(c)a_\lambda(k) =$$

$$= \frac{1}{c}\sum_{k\in\Re}a_\lambda(k)\frac{\partial}{\partial\omega_\mu}\mathbf{q}_k(c) = \frac{1}{c^2}\sum_{k\in R}a_\lambda(k)\mathbf{q}_k(c)\left(a_\mu(k)-\bar{a}_\mu\right) =$$

$$= \frac{1}{c^2}\left(\sum_{k\in R}\mathbf{q}_k(c)a_\lambda(k)a_\mu(k) - \sum_{k\in R}\mathbf{q}_k(c)a_\lambda(k)\bar{a}_\mu\right) = \frac{1}{c^2}\left(\sum_{k\in R}\mathbf{q}_k(c)a_\lambda(k)a_\mu(k) - \bar{a}_\mu\bar{a}_\lambda\right)$$

$$= \frac{1}{c^2}\left(\bar{a}_{\mu\cup\lambda} - \bar{a}_\mu\bar{a}_\lambda\right)$$

tengase en cuenta que $a_\lambda(k)a_\mu(k) = \prod_{\substack{u\in\lambda\\ v\in\mu}}k(u)k(v)$

cuando los rangos de las unidades son positivos: $R_i \subset N\cup\{0\}$ los niveles de activación son positivos y acotados

$$\left.\begin{array}{l} 0 \le a_\lambda(k)a_\mu(k) \le \left(\max_i\left\{\max(R_i)\right\}\right)^{|\lambda|+|\mu|} \\ 0 \le \bar{a}_\lambda\bar{a}_\mu < \left(\max_i\left\{\max(R_i)\right\}\right)^{|\lambda|+|\mu|} \end{array}\right\} \Rightarrow \frac{\partial^2 D(\mathbf{q}/\mathbf{q}')}{\partial\omega_\mu\partial\omega_\lambda} \le \frac{1}{c^2}\left(\max_i\left\{\max(R_i)\right\}\right)^{|\lambda|+|\mu|}$$

Se puede demostrar a partir de la acotación de las derivadas segundas que la cota superior que garantiza la convergencia del descenso por el gradiente es de la forma:

$$\alpha \le \frac{c^2}{|L|^2\left(\max_i\left\{\max(R_i)\right\}\right)^{\max(|\lambda|+|\mu|)}}$$