ESANN'1994 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 20-21-22 April 1994, D-Facto public., ISBN 2-9600049-1-4, pp. 117-122

High-order Boltzmann Machines applied to the Monk´s problems

M. Graña, V. Lavin, A. D´Anjou, F.X. Albizuri, J.A. Lozano

Dept. CCIA, UPV/EHU+ Apartado 649, 20080 San Sebastián
e-mail: ccpgrrom@si.ehu.es

Abstract: This work presents the empirical results of the application of High-order Boltzmann Machines (HOBM) to the so-called Monk´s problems. High -order Boltzmann Machines with discrete state units (non-binary units) are also introduced and applied.

0 Introduction

The Monk´s problems [1] have been used as a benchmark for the evaluation of a bread spectrum of Machine Learning algorithms, mostly algorithms for the construction of classification trees, but also some neural network algorithms (at least Backpropagation and Cascade Correlation) have been reported. This work can be assumed as an extension of [1] to include High-order Boltzmann Machines (HOBM) (with binary and discrete state units) to this comparative report. All the experiments were performed upon the original data obtained via remote ftp from the public directory referred in [1].
High-order neural networks are receiving more attention in recent times [9]. Classical references of Boltzmann Machines (BM) are [6,7]. The first hint of the possibility of defining HOBM was in [8], but up to now little attention has been paid to them. In [2] we have shown that HOBM with binary (0,1) units (1) can be trained with the same algorithm than conventional (order 2) BM, (2) if they are completely connected they can arbitrarily approximate any probability distribution in the space$\{0,1\}^N$., and (3) if not completely connected the order of the connections determines the degree of fitness that can be obtained. Besides that, in [3,4] we have explored the application of the BM to the resolution of the SAT problem. It is clear work that high order connections with 0-1 extreme units can be interpreted as AND clauses. This interpretation allows in the present work to define "a priori" topologies fitted for each of learning problems, deduced from their logical statement. Obviously, these "a priori" topologies are usually unknown, being one of the tasks of learning algorithms to uncover them, but they serve two purposes in our experiments: (1) to evaluate the Weight adaptation algorithm without "topological interferences" and (2) to evaluate the ability of the learning algorithm to uncover them . For non-binary HOBM the "a priori" topologies correspond to structures that give the appropriate maxima of the consensus function. (We perform consensus maximisation as in [7]). Section 1 introduces the Monk´s problems. Section 2 shows the results of applying binary HOBM to them. Section 3 introduces HOBM with discrete state units. Section 4

+ Work supported by the Dept. Educación, Univ. e Inv. of the Gobierno Vasco, project GV9220

ESANN'1994 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 20-21-22 April 1994, D-Facto public., ISBN 2-9600049-1-4, pp. 117-122

shows the results of their application to the Monk´s problems. Finally, section 5 gives some conclusions and further work

1 The Monk´s problems

The Monk´s problems were defined in [1] over an artificial robot domain, where each robot is described by six discrete variables:

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

Each learning problem is defined by a logical expression involving those variables, that defines the class of robots that must be discovered by the learning algorithms. (Monk´s problems are two class problems). Training and test data are produced following the logical definitions. The test data are the class assignment to the whole space (432 feature vectors), train data are random subsets of the test data. The methodology used in [1] consists in the elaboration of the model using the train data and testing it against the test data. The results reported for each learning algorithm are the percentage of correct answers to the test set. The logical definition of each problem follows:

M_1 is defined by the relation: *(head_shape = body_shape) or (jacket_color = red)*
M_2 is defined by the relation: *Exactly two of the six attributes have their first value*
M_3 is defined by the relation: *(jacket_color is green)and holding a sword) or (jacket_color is not blue and body_shape is not octagon).*

M_1 is a normal disjunctive form, and it is supposed to be easily learned by any symbolic algorithm. M_2 is close to a parity problem, difficult to state as a DNF or CNF. Finally, M_3 contains a 5% noisy data in the training set, and is intended to evaluate the robustness in the presence of noise.

2 Results with binary HOBM

First we will introduce the notation used: A HOBM is described by the triplet (U,L,W), being U the set of units and L the set of connections. Each connection $\lambda \in L$ is a subset of U. The order of a connection is its cardinal, and the order of the HOBM is that of its higher order connection. W are the weights associated to the connections, and can be formulated as $W:L \rightarrow R$. The consensus function to be maximised by the HOBM, where k(u) is the state of unit u in the global configuration k, is.

$$C(k) = \sum_{\lambda} \omega_{\lambda} \prod_{u \in \lambda} k(u)$$

In our application to the Monk´s problems the set of units is built up as follows:

ESANN'1994 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 20-21-22 April 1994, D-Facto public., ISBN 2-9600049-1-4, pp. 117-122

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

$$U_1 = \{u_{ij} | i \in \{1,2,4\}, j \in \{1..3\}\} \quad U_2 = \{u_{ij} | i \in \{3,6\}, j = 1\}$$

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

For the problem variables x_1,x_2,x_4,x_5 whose ranges are non-binary, the unit u_{ij} takes its state 1 when variable x_i takes the j-th value of its range. For the binary variables x_3 and x_6 a single unit is used. Unit u_o gives the classification output. The "a priori" topologies, deduced from the logical interpretation of connections as AND clauses, are as follows. For the M_1 problem:

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

$$L^a_{M1} = \{\{u_{1j}, u_{2j}, u_o\} j = 1..3\} \cup \{\{u_{5,1}, u_o\}\} \qquad L^i_{M1} = \{\{u_o\}\}$$

We can "a priori" distinguish L^a_{M1} (the set of excitatory connections) from L^i_{M1} (the set of inhibitory connections). A complete "a priori" HOBM for the M_1 problem would be specified by: (U^{16}, L_{M1}, W_{M1}), where

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

Similar "a priori" topologies and weights can be deduced for the M_2 and M_3 problems. Lack of space prevents their specification here. In the experiments that follow, we call densely connected HOBM of order *r* the ones with connections up to order *r* from the input units to the output unit. Formally:

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

The weight adaptation algorithm used is: $\Delta\omega_\lambda = \begin{cases} 1 & (p'_\lambda - p_\lambda) > 0 \\ -1 & (p'_\lambda - p_\lambda) < 0 \end{cases}$ Where p'_λ is the activation probability of the connection λ in the clamped phase and p_λ in the free phase. Inital weights are zero in all cases. We have considered that the input units are clamped in the free phase, this provides a better convergence of the learning algorithm, and a monotone decrease of the quadratic error $\varepsilon^2 = \sum_\lambda (p'_\lambda - p_\lambda)^2$.

We have tested two topological design schemes: pruning and weight decay. Pruning consists in the elimination of those connections for which the estimated standard deviation of the weight is bigger that its estimated mean. These estimations are computed as the weight adaptation proceeds, and the pruning is activated when the quadratic error goes below an specified threshold. Weight decay is performed as usual [5]:$\Delta\omega^d_\lambda = \Delta\omega_\lambda - \theta\omega_\lambda$ with a pruning at the end of the learning algorithm of connections whose weight absolute magnitude is less than 1.

Finally, for the sake of completeness we have performed experiments with conventional BM with 3,2 and 4 hidden units for the M_1, M_2 and M_3 problems respectively. Table 1 shows the results of our experiments.

ESANN'1994 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 20-21-22 April 1994, D-Facto public., ISBN 2-9600049-1-4, pp. 117-122

Rows of table 1 show (from top to bottom) the best result reported in [1] (when #con is specified the winner was backpropagation), the application of the a priori HOBM with the weights specified above, the results of the weight adaptation upon the a priori topology, and the results of application to densely connected topologies L^r, increasing the order, of the raw weight adaptation algorithm, the pruning algorithm activated at different thresholds and weight decay. The experiments have proceed increasing the order of the densely connected topology, searching for better results. In the case of M_1 only order 3 was needed, in the case of M_2 bad results were obtained regardless of the order, and for M_3 the increased order improves the learning ability of the machine. In general it can be observed that the lower the activation threshold the better results of the pruning algorithm, sometimes improving over the unpruned topology. Weight decay performs similar to the proposed pruning.

	M1		M2		M3	
	%hits	#con	%hits	#con	%hit	#con
Best result in [1]	100	58	100	41	100	-
A priori HOBM	100	5	99,8	36	100	4
A priori topology	100	5	96,75	36	97,22	4
L^3	100	106	60	106	92.43	106
pruning $\varepsilon^2 < 0.1$	100	17	51,8	0	52.77	1
0.01	100	22	-		96.99	33
weight decay θ=0.1			-		95,37	29
L^4	-		72.68	380	93,75	380
pruning $\varepsilon^2 < 0.1$	-		67.12	76	51,8	1
0.01	-		72.45	145	94.9	128
weight decay θ=0.1			58,56	380	93,98	130
L^5	-		71.99	821	93.25	821
pruning $\varepsilon^2 < 0.1$	-		64.81	255	47.22	1
0.01	-		68.05	313	86.80	89
weight decay θ=0.1			68,98	326	94,67	93
L^6	-		70,86	1172	96,75	1172
pruning $\varepsilon^2 < 0.1$	-		65,7	264	91,2	58
weight decay θ=0.1	-		70,3	804	93,28	774
L^7	-		71,9	1280	96,29	1280
Hidden units	91,3	55	67,12	36	96,75	75

Table 1. Results with binary HOBM

3 Non binary HOBM

Our notation for the discrete HOBM is an straightforward extension of the binary case. A HOBM with discrete state units is described by the cuadrupla (U,R,L,W) where U, L, W preserve their meaning and $R = \{R_i \subset \mathbb{Z}\}$ where R_i is the range of values of the state of unit u_i. The multiplicative interpretation of the connections is

ESANN'1994 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 20-21-22 April 1994, D-Facto public., ISBN 2-9600049-1-4, pp. 117-122

preserved, so the consensus function preserves its general form: $C(k)=\sum_{\lambda}\omega_{\lambda}\prod_{u\in\lambda}k(u)$ where $k(u_i)\in R_i$.. It can be show that for an observable probability distribution **q'** to be feasible by a discrete state HOBM, a necessary condition is that 0 belongs to the state range of all units, formally: $\forall R_i \in R, 0 \in R_i$.

The weight adaptation rule can be deduced as a gradient descent of the information divergence measure, the so-called Kullback distance, D(**q'**/**q**) between the desired distribution **q'** the one modelled by the HOBM **q**. This gradient is of the form: $\frac{\partial D(q'/q)}{\partial \omega_{\lambda}} = -\frac{1}{c}(a'_{\lambda}-a_{\lambda})$ where $a_{\lambda} = \sum_{k} q_k \prod_{u\in\lambda} k(u)$ is the mean activation level of connection λ under a distribution $\{q_k\}$ of the global configurations. Convergence conditions similar to the ones deduced in [2] for the binary HOBM can also be obtained in this case, taking into account the second derivatives of D relative to the weights.

4 Results of the discrete state HOBM

The set of units considered for the Monk´s problems is $U=\{u_i\ i=1..6, u_o\}$, the state ranges of the units are: $R_1=R_2=R_4=\{0..2\}$, $R_3=R_6=R_o=\{0..1\}$ and $R_5=\{0..3\}$. A priori topologies can be deduced taking into consideration the possible weight relations that could give the consensus function maxima for the selected global configurations. This task seems impossible for M_1 y M_3, however, it is possible for M_2. Lack of space prevents the explicit formulation of the "a priori" topology and weights. Densely connected HOBM of order *r* have topologies $L^r=\{\lambda\subset U | (|\lambda|\le r)\wedge(u_o\in\lambda)\}$

Table 2 shows the results of the application of discrete state HOBM to the Monk´s problems, the interpretation of the rows is the same as in table 1.

	M1		M 2		M3	
	%hits	#con	%hits	#con	%hit	#con
Best result in [1]	100	58	100	41	100	-
A priori HOBM	-		100	22	-	
A priori topology	-		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
L^6	81	63	91,9	63	91,8	63

Table 2. Results with non-binary HOBM

ESANN'1994 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 20-21-22 April 1994, D-Facto public., ISBN 2-9600049-1-4, pp. 117-122

5 Conclusions and further work

In relation with the binary HOBM: The AND interpretation of the connections allows concise topological formulation well fitted to the problem. When an "a priori" topology can defined the convergence to the desired probability distribution is guaranteed. This means that good intuitions of the topology could give very good results, and it seems easier to have these good intuitions reasoning in terms of high order connections. However, in the very common case when no good hints of the logical structure of the problem can be obtained, still there are no good schemes to uncover this structure, that is, we have no efficient pruning algorithms.
The discrete state HOBM provide, when applicable, provide extremely compact formulations. The class of distributions that can be modelled seems to be more restricted than in the binary case, because of the collapse of many states that become indistinguishable. However, the learning algorithm seems to be able to obtain good approximations to the desired distributions. Also in this case the tested pruning algorithms seem of little use.
We are driving our efforts to continue the application of HOBM to other learning problems for which public databases are available. Also we continue to search for pruning schemes able to discover the appropriate topologies. Finally we are interested in finding a characterisation of the distributions that the HOBM with discrete states is able to fit, in order to determine its applicability to other problems, such as sound and image compression.

References

[1] Thrun S.B. et alters "The MONK´s problems: A performance comparison of different learning algorithms" Report CMU-CS-91-197 Carnegie Melon Univ.
[2] F.X. Albizuri, A. D´Anjou, M. Graña, F.J. Torrealdea, M.C. Hernandez "The High Order Boltzmann Machine: learned distribution and topology" IEEE Trans. Neural Networks in press
[3] A. D´Anjou, 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
[4] A. D´Anjou, M. Graña, F.J. Torrealdea, M.C. Hernandez "Solving satisfiability via Boltzmann Machines" IEEE Trans. on Patt. An. and Mach. Int. Mayo 93
[5] G.E. Hinton, Lectures at the Neural Network Summer School, Wolfson College, Cambridge, Sept. 1993
[6] D.H. Ackley, G.E. Hinton, T.J. Sejnowski "A learning algorithm for Boltzmann Machines" Cogn. Sci. 9 (1985) pp.147-169
[7] E.H.L. Aarts, J.H.M. Korst "Simulated Annealing and Boltzmann Machines: a stochastic approach to combinstorial optimization and neural computing" John Wiley & Sons (1989)
[8] TJ. Sejnowski "Higher order Boltzmann Machines" in Denker (ed) Neural Networks for computing AIP conf. Proc. 151, Snowbird UT (1986) pp.398-403
[9] S.J. Perantonis, 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