The Connectionist Inductive Learning and Logic Programming System

Artur S. d'Avila Garcez
Department of Computing - Imperial College
180 Queen's Gate, SW7- 2BZ, London - UK
aag@doc.ic.ac.uk

Gerson Zaverucha
COPPE Engenharia de Sistemas e Computação - UFRJ
Caixa Postal: 68511, CEP: 21945-970, Rio de Janeiro - Brazil
gerson@cos.ufrj.br

November 17, 1998

Abstract

This paper presents the Connectionist Inductive Learning and Logic Programming System (*C-IL^{2}P*). *C-IL^{2}P* is a new massively parallel computational model based on a feedforward Artificial Neural Network that integrates inductive learning from examples and background knowledge with deductive learning from Logic Programming. Starting with the background knowledge represented by a propositional logic program, a translation algorithm is applied generating a neural network that can be trained with examples. The results obtained with this refined network can be explained by extracting a revised logic program from it. Moreover, the neural network computes the stable model of the logic program inserted in it or learned with the examples, as a parallel system for Logic Programming. We have successfully applied *C-IL^{2}P* in two real-world problems on computational biology, specifically DNA sequence analyses. Comparisons with the results obtained by some of the main neural, symbolic, and hybrid inductive learning systems in the same domain knowledge show the effectiveness of *C-IL^{2}P*.

Keywords: Theory Refinement, Machine Learning, Artificial Neural Networks, Logic Programming, Computational Biology.

1 Introduction

The aim of neural-symbolic integration is to explore the advantages that each paradigm presents. Within the features of artificial neural networks are its massive parallelism, inductive learning and generalization capabilities. On

the other hand, symbolic systems can explain their inference process (e.g. through automatic theorem proving), and use powerful declarative languages for knowledge representation.

"It is generally accepted that one of the main problems in building Expert Systems (which are responsible for the industrial success of Artificial Intelligence) lies in the process of knowledge acquisition, known as the *knowledge acquisition bottleneck*" [29]. An alternative is the automation of this process using Machine Learning techniques [35]. Symbolic machine learning methods are usually more effective if they can exploit a background knowledge (incomplete domain theory). In contrast, neural networks have been successfully applied as a learning method from examples only (data learning) [53]. As a result, the integration of theory and data learning in neural networks seems to be a natural way towards more powerful training mechanisms.

Learning strategies can be classified as: learning from instruction, learning by deduction, learning by analogy, learning from examples and learning by observation and discovery [33]; the latter two are forms of inductive learning. The inductive learning task is to find hypotheses that are consistent with a background knowledge to explain a given set of examples. In general, those hypotheses are definitions of concepts described in some logical language, the examples are descriptions of instances and non-instances of the concept to be learned, and the background knowledge provides additional information about the examples and the concepts' domain knowledge [29].

In contrast to symbolic learning systems, neural networks' learning implicitly encodes patterns and their generalizations in the networks' weights, which reflect the statistical properties of the trained data [9]. It has been indicated that neural networks can outperform symbolic learning systems, specially when data are noisy [53]. This result, that is also due to the massively parallel architecture of neural networks, contributed decisively to the growing interest on combining, and possibly integrating, neural and symbolic learning systems (see [28] for a clarifying treatment on the suitability of neural networks for the representation of symbolic knowledge).

Pinkas [41, 42] and Holldobler [22] have given important contributions for the subject of neural-symbolic integration, showing the capabilities and limitations of neural networks at performing logical inference. Pinkas defined a bi-directional mapping between symmetric neural networks and mathematical logic [12]. He presented a theorem showing a weak equivalence between the problem of satisfiability of propositional logic and minimizing energy functions. The equivalence is in the sense that for every *well-formed formula (wff)* a quadratic energy function can efficiently be found and for every energy function there exists a *wff* (inefficiently found) such that the global minima of the function are exactly equal to the satisfying models of the formula. Holldobler presented a parallel unification algorithm and an automated reasoning system for first order Horn clauses, implemented

in a feedforward neural network, called *Connectionist Horn Clause Logic (CHCL) System*.

In [23], Holldobler and Kalinke presented a method for inserting propositional general logic programs [30] into three-layer feedforward artificial neural networks. They have shown that for each program $\mathcal{P}$ there exists a three-layer feedforward neural network $\mathcal{N}$ with binary threshold units that computes $T_{\mathcal{P}}$, the program's fixed point operator. If $\mathcal{N}$ is transformed into a recurrent network by linking the units in the output layer to the corresponding units in the input layer, it always settles down in a unique stable state when $\mathcal{P}$ is an acceptable program[1] [4, 14]. This stable state is the least fixed point of $T_{\mathcal{P}}$, that is identical to the unique stable model of $\mathcal{P}$, what provides a declarative semantics for $\mathcal{P}$ (see [18] for the stable model semantics of general logic programs).

In [58], Towell and Shavlik presented *KBANN (Knowledge-based Artificial Neural Network)*, a system for rules' insertion, refinement and extraction from neural networks. They have empirically shown that knowledge-based neural networks' training based on the backpropagation learning algorithm [48] is a very efficient way to learn from examples and background knowledge. They have done that by comparing KBANN's performance with other hybrid, neural and purely symbolic inductive learning systems' (see [29] and [36] for a comprehensive description of symbolic inductive learning systems, including Inductive Logic Programming).

The *Connectionist Inductive Learning and Logic Programming* (*C-IL^{2}P*) system is a massively parallel computational model based on a feedforward artificial neural network that integrates inductive learning from examples and background knowledge with deductive learning from Logic Programming. Starting with the background knowledge represented by a propositional general logic program, a translation algorithm is applied (see figure 1 (1)) generating a neural network that can be trained with examples (2). Furthermore, that neural network computes the stable model of the program inserted in it or learned with the examples as a massively parallel system for Logic Programming (3). The result of refining the background knowledge with the training examples can be explained by extracting a revised logic program from the network (4). Finally, the knowledge extracted can be more easily analyzed by an expert that decides if it should feed the system once more, closing the learning cycle (5).

In section 2, we present a new translation algorithm from general logic programs ($\mathcal{P}$) to artificial neural networks ($\mathcal{N}$) with bipolar semi-linear neurons. The algorithm is based on Holldobler and Kalinke's translation algorithm from general logic programs to neural networks with binary threshold neurons [23]. We also present a theorem showing that $\mathcal{N}$ computes the fixed-point operator ($T_{\mathcal{P}}$) of $\mathcal{P}$. The theorem ensures that our translation

[1] An acceptable program P has exactly one stable model.

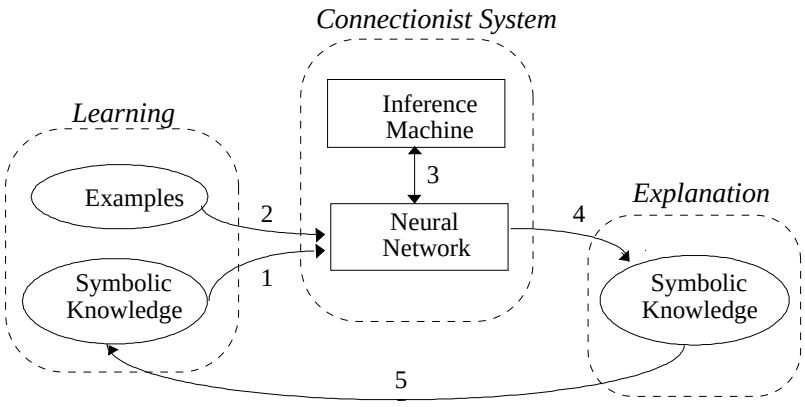

Figure 1: The Connectionist Inductive Learning and Logic Programming System.

algorithm is sound. In section 3, we show that the result obtained in [23] still holds, that is, $\mathcal{N}$ is a massively parallel model for Logic Programming. However, $\mathcal{N}$ can also perform inductive learning from examples efficiently, assuming $\mathcal{P}$ as background knowledge and using the standard backpropagation learning algorithm as in [58]. We outline the steps for performing both deduction and induction in the neural network. In section 4, we successfully apply the system in two real-world problems of DNA classification, which have become benchmark data sets for testing machine learning systems' accuracy. We compare the results with other neural, symbolic, and hybrid inductive learning systems. Briefly, *C-IL^{2}P's* test-set performance is at least as good as *KBANN's* and better than any other system investigated, while its training-set performance is considerably better than *KBANN's*. In section 5, we conclude and discuss directions for future work.

2 Translation From Logic Programs to Neural Networks

We have seen that the merging of theory (background knowledge) and data learning (learning from examples) in neural networks may provide a more effective learning system. Towards that goal, one could firstly translate the background knowledge into a neural network initial architecture, and then train it with examples using some neural learning algorithm like backpropagation. In order to do so, the *C-IL^{2}P* system provides a translation algorithm from propositional (or grounded) general logic programs to feedforward neural networks with semi-linear neurons, and a theorem showing

that the network obtained is equivalent to the original program, in the sense that what is computed by the program is computed by the network and vice-versa.

Definition 1 *A general clause is a rule of the form* $A \leftarrow L_1, ..., L_k$, *where* A *is an atom and* L_i $(1 \leq i \leq k)$ *is a literal (an atom or the negation of an atom). A general logic program is a finite set of general clauses.*

To insert the background knowledge, described by a general logic program ($\mathcal{P}$), in the neural network ($\mathcal{N}$), we use an approach similar to Holldobler and Kalinke's [23]. Each general clause (C_l) of $\mathcal{P}$ is mapped from the input layer to the output layer of $\mathcal{N}$ through one neuron (N_l) in the single hidden layer of $\mathcal{N}$. Intuitively, the translation algorithm from $\mathcal{P}$ to $\mathcal{N}$ has to implement the following conditions: *(1)* The input potential of a hidden neuron (N_l) can only exceed N_l's threshold (θ_l), activating N_l, when all the positive antecedents of C_l are assigned the truth-value "$true$" while all the negative antecedents of C_l are assigned "$false$"; and *(2)* The input potential of an output neuron (A) can only exceed A's threshold (θ_A), activating A, when at least one hidden neuron N_l that is connected to A is activated.

Example 2 *Consider the logic program* $\mathcal{P} = \{A \leftarrow B, C, notD;\ A \leftarrow E, F;\ B \leftarrow\}$. *The translation algorithm should derive the network* $\mathcal{N}$ *of figure 2, setting weights* ($W's$) *and thresholds* ($\theta's$) *in such a way that conditions* (1) *and* (2) *above are satisfied. Note that, if* $\mathcal{N}$ *ought to be fully-connected, any other link (not shown in figure 2) should receive weight zero initially.*

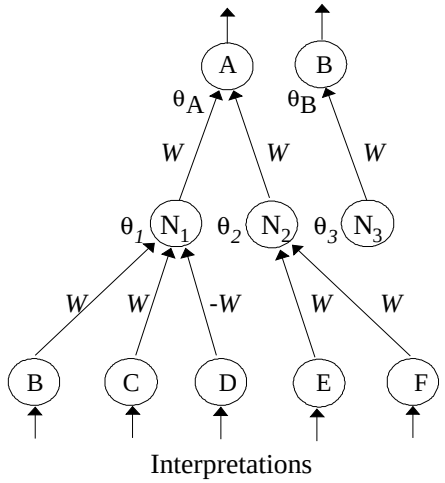

Figure 2: Sketch of a neural network for the above logic program P.

Note that, in example 2, we have labelled each input and output neuron as an atom appearing, respectively, in the body and in the head of a clause of $\mathcal{P}$. That allows us to refer to neurons and propositional variables interchangeably and to regard each network's input vector $\mathbf{i} = (i_1, ..., i_m)$ ($i_{j(1 \leq j \leq m)} \in [-1, 1]$) as an interpretation for $\mathcal{P}$, where if $i_j \in [A_{min}, 1]$ then the propositional variable associated to the j-th neuron in the network's input layer is assigned "$true$", while $i_j \in [-1, -A_{min}]$ means that it is assigned "$false$". Note also that each hidden neuron N_l corresponds to a clause C_l of $\mathcal{P}$. $A_{min} \in (0, 1)$ is a predefined value, as will be seen in the sequel.[2]

The following notation will be used in our translation algorithm.

Notation : Given a general logic program $\mathcal{P}$, let q denote the number of clauses C_l $(1 \leq l \leq q)$ occurring in $\mathcal{P}$;

m, the number of literals occurring in $\mathcal{P}$;

A_{min}, the minimum activation for a neuron to be considered "active" (or "$true$"), $0 < A_{min} < 1$;

A_{max}, the maximum activation for a neuron to be considered "not active" (or "$false$"), $-1 < A_{max} < 0$;

$h(x) = \frac{2}{1+e^{-\beta x}} - 1$, the bipolar semi-linear activation function, where β is the steepness parameter (that defines the slope of $h(x)$);

$g(x) = x$, the standard linear activation function;

W (resp. -W), the weight of connections associated with positive (resp. negative) literals;

θ_l, the threshold of hidden neuron N_l associated with clause C_l;

θ_A, the threshold of output neuron A, where A is the head of clause C_l;

k_l, the number of literals in the body of clause C_l;

p_l, the number of positive literals in the body of clause C_l;

n_l, the number of negative literals in the body of clause C_l;

μ_l, the number of clauses in $\mathcal{P}$ with the same atom in the head for each clause C_l;

$MAX_{C_l}(k_l, \mu_l)$, the greatest element between k_l and μ_l for clause C_l,

$MAX_{\mathcal{P}}(k_1, ..., k_q, \mu_1, ..., \mu_q)$, the greatest element among all k's and μ's of $\mathcal{P}$.

For instance, for the program $\mathcal{P}$ of example 2 $q = 3$, $m = 6$, $k_1 = 3$, $k_2 = 2$, $k_3 = 0$, $p_1 = 2$, $p_2 = 2$, $p_3 = 0$, $n_1 = 1$, $n_2 = 0$, $n_3 = 0$, $\mu_1 = 2$, $\mu_2 = 2$,

[2] An *interpretation* is a function from propositional variables to {"$true$", "$false$"}. A *model* for P is an interpretation that maps P to "$true$".

$\mu_3 = 1$, $MAX_{C_1}(k_1, \mu_1) = 3$, $MAX_{C_2}(k_2, \mu_2) = 2$, $MAX_{C_3}(k_3, \mu_3) = 1$, and $MAX_{\mathcal{P}}(k_1, k_2, k_3, \mu_1, \mu_2, \mu_3) = 3$.

In the translation algorithm below, we define $A_{min} = \xi_1(k, \mu)$, $W = \xi_2(h(x), k, \mu, A_{min})$, $\theta_l = \xi_3(k, A_{min}, W)$, *and* $\theta_A = \xi_4(\mu, A_{min}, W)$ such that conditions *(1)* and *(2)* are satisfied, as we will see later in the proof of theorem 3.

Given a general logic program $\mathcal{P}$, consider that the literals of $\mathcal{P}$ are numbered from 1 to m such that the input and output layers of $\mathcal{N}$ are vectors of maximum length m, where the i-th neuron represents the i-th literal of $\mathcal{P}$. Assume, for mathematical convenience and without loss of generality, that $A_{max} = -A_{min}$.

1. Calculate $MAX_{\mathcal{P}}(k_1, ..., k_q, \mu_1, ..., \mu_q)$ of $\mathcal{P}$;
2. Calculate $A_{min} > \frac{MAX_{\mathcal{P}}(k_1,...,k_q,\mu_1,...,\mu_q)-1}{MAX_{\mathcal{P}}(k_1,...,k_q,\mu_1,...,\mu_q)+1}$;
3. Calculate $W \geq \frac{2}{\beta} \cdot \frac{ln(1+A_{\min})-ln(1-A_{\min})}{MAX_{\mathcal{P}}(k_1,...,k_q,\mu_1,...,\mu_q)(A_{\min}-1)+A_{\min}+1}$;
4. For each clause C_l of $\mathcal{P}$ of the form $A \leftarrow L_1, ..., L_k$ $(k \geq 0)$:
   (a) Add a neuron N_l to the hidden layer of $\mathcal{N}$;
   (b) Connect each neuron L_i $(1 \leq i \leq k)$ in the input layer to the neuron N_l in the hidden layer. If L_i is a positive literal then set the connection weight to W; otherwise, set the connection weight to $-W$;
   (c) Connect the neuron N_l in the hidden layer to the neuron A in the output layer and set the connection weight to W;
   (d) Define the threshold (θ_l) of the neuron N_l in the hidden layer as $\theta_l = \frac{(1+A_{\min})(k_l-1)}{2}W$.
   (e) Define the threshold (θ_A) of the neuron A in the output layer as $\theta_A = \frac{(1+A_{\min})(1-\mu_l)}{2}W$.
5. Set $g(x)$ as the activation function of the neurons in the input layer of $\mathcal{N}$. In this way, the activation of the neurons in the input layer of $\mathcal{N}$, given by each input vector $\mathbf{i}$, will represent a interpretation for $\mathcal{P}$.
6. Set $h(x)$ as the activation function of the neurons in the hidden and output layers of $\mathcal{N}$. In this way, a gradient descent learning algorithm, such as *backpropagation*, can be applied on $\mathcal{N}$ efficiently.
7. If $\mathcal{N}$ ought to be fully-connected, set all other connections to zero.

Since $\mathcal{N}$ contains a bipolar semi-linear (differentiable) activation function $h(x)$, instead of a binary threshold non-linear activation function, the network's output neurons activations are real numbers in the range $[-1, 1]$. Therefore, we define that an output within the range $[A_{min}, 1]$ represents the truth-value "$true$", while an output within $[-1, -A_{min}]$ represents "$false$". We will see later in the proof of theorem 3 that the above defined weights and thresholds do not allow the network to present activations in the range $(-A_{min}, A_{min})$.

Note that the translation of facts of $\mathcal{P}$ into $\mathcal{N}$, for instance $B \leftarrow$ in example 2, is done by simply taking $k = 0$ in the above algorithm. Alternatively, each fact of the form $A \leftarrow$ may be converted to a rule of the form $A \leftarrow T$ that is inserted in $\mathcal{N}$ using $k = 1$, where T, that stands for "$true$", is an extra neuron that is always active in the input layer of $\mathcal{N}$, that is, T has input data fixed at "1". From the point of view of the computation of $\mathcal{P}$ by $\mathcal{N}$, there is absolutely no difference between the above two ways of inserting facts of $\mathcal{P}$ into $\mathcal{N}$. However, considering the subsequent process of inductive learning, looking at $\mathcal{P}$ as background knowledge, if $A \leftarrow$ is inserted in $\mathcal{N}$ then the set of examples to be learned afterwards can defeat that fact by changing weights and/or establishing new connections in $\mathcal{N}$. On the other hand, if $A \leftarrow T$ is inserted in $\mathcal{N}$ then A can not be defeated by the set of examples since the neuron T is clamped in $\mathcal{N}$. Defeasible and nondefeasible knowledge can, therefore, be inserted in the network by defining variable or fixed weights and neurons, respectively.

The above translation algorithm is based upon the one presented in [23], where $\mathcal{N}$ is defined with binary threshold neurons. It is known that such networks possess little learnability. Here, in order to perform inductive learning efficiently, $\mathcal{N}$ is defined using the activation function $h(x)$. An immediate result is that $\mathcal{N}$ can also perform inductive learning from examples and background knowledge, as in [58], using for instance the backpropagation learning algorithm. Moreover, the restriction imposed over W in [23], where it is shown that $\mathcal{N}$ computes T_P for $W = 1$, is weakened here, since the weights must be able to change during training.

Nevertheless, in [58], and more clearly in [56], the background knowledge must have a "*sufficiently small*" number of rules as well as a "*sufficiently small*" number of antecedents in each rule[3] in order to be accurately encoded in the neural network. Unfortunately, these restrictions become quite strong or even unfeasible if, for instance, $A_{max} = \frac{1}{2}$, as in ([58], Section 5: Empirical Tests of KBANN). In that way, an interpretation that does not satisfy a clause can wrongly activate a neuron in the output layer of $\mathcal{N}$. That results from the use of the standard unipolar activation function, where each neuron's activation is in the range $[0, 1]$. Consequently, both "$false$"

[3] The "*sufficiently small*" restrictions are given by equations $\mu A_{max} \leq \frac{1}{2}$ and $k A_{max} \leq \frac{1}{2}$, respectively, where $A_{max} > 0$ [56].

and "*true*" are represented by *positive* numbers, in the range $[0, A_{max}]$ and $[A_{min}, 1]$ respectively, and taking for instance $A_{min} = 0.7$ and $k = 2$, an interpretation that assigns "$false$" to positive literals in the input layer of $\mathcal{N}$ can generate a *positive* input potential greater than the hidden neuron's threshold, wrongly activating the neuron in the output layer of $\mathcal{N}$.

In order to solve this problem we use bipolar activation functions, where each neuron's activation is in the range $[-1, 1]$. Now, an interpretation that does not satisfy a clause contributes *negatively* to the hidden neuron's input potential, since "$false$" is represented by a number in $[-1, -A_{min}]$, while an interpretation that does satisfy a clause contributes *positively* to the input potential, because "*true*" is in $[A_{min}, 1]$. Theorem 3 will show that the choice of a bipolar activation function is sufficient to solve the above problem. Furthermore, the choice of -1 instead of *zero* to represent "$false$" will lead to faster convergence in almost all cases. The reason is that the update of a weight connected to an input variable will be *zero* when the corresponding variable is *zero* in the training pattern [20, 9].

Considering, then, bipolar semi-linear activation functions $h(x)$, let us see how we have obtained the values for the hidden and output neurons' thresholds θ_l and θ_A. Assume, without loss of generality, that $A_{max} = -A_{min}$, what confers symmetric mathematical results. From the input to the hidden layer of $\mathcal{N}$ $(L_1, ..., L_k \Rightarrow N_l)$, if an interpretation satisfies $L_1, ..., L_k$ then the contribution of $L_1, ..., L_k$ to the input potential of N_l is greater than $I_+ = kA_{min}W$. If, conversely, an interpretation does not satisfy $L_1, ..., L_k$ then the contribution of $L_1, ..., L_k$ to the input potential of N_l is smaller than $I_- = (p-1)W - A_{min}W + nW$. Therefore, we define $\theta_l = \frac{I_+ + I_-}{2} = \frac{(1+A_{min})(k-1)}{2}W$ (Translation Algorithm, step 4d). From the hidden to the output layer of $\mathcal{N}$ $(N_l \Rightarrow A)$, if an interpretation satisfies N_l then the contribution of N_l to the input potential of A is greater than $I_+ = A_{min}W - (\mu - 1)W$. If, conversely, an interpretation does not satisfy N_l then the contribution of N_l to the input potential of A is smaller than $I_- = -\mu A_{min}W$. Similarly, we define $\theta_A = \frac{I_+ + I_-}{2} = \frac{(1+A_{min})(1-\mu)}{2}W$ (Translation Algorithm, step 4e). Obviously, $I_+ > I_-$ should be satisfied in both cases above. Therefore, $A_{min} > \frac{k_l - 1}{k_l + 1}$ and $A_{min} > \frac{\mu_l - 1}{\mu_l + 1}$ must be verified and, more generally, the condition imposed over A_{min} in the translation algorithm (step 2). Finally, given A_{min}, the value of W (translation algorithm (step 3)) results from the proof of theorem 3 below.

In the sequel, we will show that the theorem presented in [23], where $\mathcal{N}$ with binary threshold neurons computes the fixed point operator $T_\mathcal{P}$ of the program $\mathcal{P}$, still holds for $\mathcal{N}$ with semi-linear neurons. The following theorem ensures that our translation algorithm is sound. The function $T_\mathcal{P}$ mapping interpretations to interpretations is defined as follows. Let $\mathbf{i}$ be an interpretation and A an atom. $T_\mathcal{P}(\mathbf{i})(A) =$ "*true*" *iff there exists* $A \leftarrow L_1, \ldots, L_k$ *in* $\mathcal{P}$ *s.t.* $\bigwedge_{i=1}^{k} \mathbf{i}(L_i) =$ "*true*".

Theorem 3 *For each propositional general logic program $\mathcal{P}$, there exists a feedforward artificial neural network $\mathcal{N}$ with exactly one hidden layer and semi-linear neurons, obtained by the above "*Translation Algorithm*", such that $\mathcal{N}$ computes $T_{\mathcal{P}}$.*

Proof. *We have to show that there exists $W > 0$ such that $\mathcal{N}$ computes $T_{\mathcal{P}}$. In order to do so, we need to prove that, given an input vector $\mathbf{i}$, each neuron A in the output layer of $\mathcal{N}$ is "active" if and only if there exists a clause of $\mathcal{P}$ of the form $A \leftarrow L_1, ..., L_k$, where $L_1, ..., L_k$ are satisfied by interpretation $\mathbf{i}$. The proof takes advantage of the monotonically non-decreasing property of the bipolar semi-linear activation function $h(x)$, that allows the analysis to focus on the boundary cases. As before, we assume that $A_{max} = -A_{min}$, what confers interesting symmetric results to the proof, without loss of generality.*

($\leftarrow$) $A \geq A_{min}$ if $L_1, ..., L_k$ is satisfied by $\mathbf{i}$. Assume that the p positive literals in "$L_1, ..., L_k$" are "true", while the n negative literals in "$L_1, ..., L_k$" are "$false$". Consider the mapping from the input layer to the hidden layer of $\mathcal{N}$. The input potential (I_l) of N_l is minimum when all the neurons associated with a positive literal in "$L_1, ..., L_k$" are at A_{min}, while all the neurons associated with a negative literal in "$L_1, ..., L_k$" are at $-A_{min}$. Thus, $I_l \geq pA_{min}W + nA_{min}W - \theta_l$ and, assuming $\theta_l = \frac{(1+A_{min})(k-1)}{2}W$, $I_l \geq pA_{min}W + nA_{min}W - \frac{(1+A_{min})(k-1)}{2}W$.

If $h(I_l) \geq A_{min}$, which means $I_l \geq -\frac{1}{\beta}ln\left(\frac{1-A_{min}}{1+A_{min}}\right)$, then N_l is active. Therefore, the following equation 1 must be satisfied.

$$pA_{min}W + nA_{min}W - \frac{(1+A_{min})(k-1)}{2}W \geq -\frac{1}{\beta}ln\left(\frac{1-A_{min}}{1+A_{min}}\right) \quad (1)$$

Solving equation 1 for the connection weight (W) yields equations 2 and 3, given that $W > 0$.

$$W \geq -\frac{2}{\beta} \cdot \frac{ln(1-A_{min}) - ln(1+A_{min})}{k(A_{min}-1) + A_{min} + 1} \quad (2)$$

$$A_{min} \rangle \frac{k-1}{k+1} \quad (3)$$

Consider, now, the mapping from the hidden layer to the output layer of $\mathcal{N}$. By equations 2 and 3, at least one neuron N_l that is connected to A is "active". The input potential (I_l) of A is minimum when N_l is at A_{min}, while the other $\mu - 1$ neurons connected to A are at -1. Thus, $I_l \geq A_{min}W - (\mu-1)W - \theta_l$ and, assuming $\theta_l = \frac{(1+A_{min})(1-\mu)}{2}W$, $I_l \geq A_{min}W - (\mu-1)W - \frac{(1+A_{min})(1-\mu)}{2}W$.

If $h(I_l) \geq A_{min}$, which means $I_l \geq -\frac{1}{\beta}ln\left(\frac{1-A_{min}}{1+A_{min}}\right)$, then A is active. Therefore, the following equation 4 must be satisfied.

$$A_{min}W - (\mu-1)W - \frac{(1+A_{min})(1-\mu)}{2}W \geq -\frac{1}{\beta}ln\left(\frac{1-A_{min}}{1+A_{min}}\right) \quad (4)$$

Solving equation 4 for the connection weight W yields equations 5 and 6, given that $W > 0$.

$$W \geq -\frac{2}{\beta} \cdot \frac{ln\,(1 - A_{min}) - ln\,(1 + A_{min})}{\mu\,(A_{min} - 1) + A_{min} + 1} \tag{5}$$

$$A_{min} \rangle \frac{\mu - 1}{\mu + 1} \tag{6}$$

($\rightarrow$) $A \leq -A_{min}$ if $L_1, ..., L_k$ is not satisfied by $\mathbf{i}$*. Assume that at least one of the p positive literals in "$L_1, ..., L_k$" is "$false$" or one of the n negative literals in "$L_1, ..., L_k$" is "$true$". Consider the mapping from the input layer to the hidden layer of $\mathcal{N}$. The input potential (I_l) of N_l is maximum when only one neuron associated to a positive literal in "$L_1, ..., L_k$" is at $-A_{min}$ or when only one neuron associated to a negative literal in "$L_1, ..., L_k$" is at A_{min}. Thus, $I_l \leq (p-1)\,W - A_{min}W + nW - \theta_l$ or $I_l \leq (n-1)\,W - A_{min}W + pW - \theta_l$, respectively, and, assuming $\theta_l = \frac{(1+A_{min})(k-1)}{2}W$, $I_l \leq (k - 1 - A_{min})\,W - \frac{(1+A_{min})(k-1)}{2}W$.*

If $-A_{min} \geq h(I_l)$, that is $-A_{min} \geq \frac{2}{1+e^{-\beta(I_l)}} - 1$, which means $I_l \leq -\frac{1}{\beta}ln\left(\frac{1+A_{min}}{1-A_{min}}\right)$ then N_l is not active. Therefore, the following equation 7 must be satisfied.

$$(k - 1 - A_{min})W - \frac{(1 + A_{min})(k-1)}{2}W \leq -\frac{1}{\beta}ln\left(\frac{1 - A_{min}}{1 + A_{min}}\right) \tag{7}$$

Solving equation 7 for the connection weight W yields equations 8 and 9, given that $W > 0$.

$$W \geq \frac{2}{\beta} \cdot \frac{ln\,(1 + A_{min}) - ln\,(1 - A_{min})}{k\,(A_{min} - 1) + A_{min} + 1} \tag{8}$$

$$A_{min} \rangle \frac{k - 1}{k + 1} \tag{9}$$

Consider, now, the mapping from the hidden layer to the output layer of $\mathcal{N}$. By equations 8 and 9, all the neurons N_l that are connected to A are "not active". The input potential (I_l) of A is maximum when all the neurons connected to A are at $-A_{min}$. Thus, $I_l \leq -\mu A_{min}W - \theta_l$ and, assuming $\theta_l = \frac{(1+A_{min})(1-\mu)}{2}W$, $I_l \leq -\mu A_{min}W - \frac{(1+A_{min})(1-\mu)}{2}W$.

If $-A_{min} \geq h(I_l)$, that is $-A_{min} \geq \frac{2}{1+e^{-\beta(I_l)}} - 1$, which means $I_l \leq -\frac{1}{\beta}ln\left(\frac{1+A_{min}}{1-A_{min}}\right)$, then A is not active. Therefore, the following equation 10 must be satisfied.

$$-\mu A_{min}W - \frac{(1 + A_{min})\,(1 - \mu)}{2}W \leq -\frac{1}{\beta}ln\left(\frac{1 + A_{min}}{1 - A_{min}}\right) \tag{10}$$

Solving equation 10 for the connection weight W yields equations 11 and 12, given that $W > 0$.

$$W \geq \frac{2}{\beta} \cdot \frac{ln\,(1 + A_{min}) - ln\,(1 - A_{min})}{\mu\,(A_{min} - 1) + A_{min} + 1} \tag{11}$$

$$A_{min} \rangle \frac{\mu - 1}{\mu + 1} \tag{12}$$

Notice that equations 2 and 5 are equivalent to equations 8 and 11, respectively. Hence, the above theorem holds if for each clause C_l in $\mathcal{P}$ equations 2 and 3 are satisfied by W and A_{min} from the input to the hidden layer of $\mathcal{N}$, while equations 5 and 6 are satisfied by W and A_{min} from the hidden to the output layer of $\mathcal{N}$.

In order to unify the weights in $\mathcal{N}$ for each clause C_l of $\mathcal{P}$, given the definition of $MAX_{C_l}(k_l, \mu_l)$, it is sufficient that equations 13 and 14 below are satisfied by W and A_{min}, respectively.

$$W \geq \frac{2}{\beta} \cdot \frac{ln\,(1 + A_{min}) - ln\,(1 - A_{min})}{MAX_{C_l}(k_l, \mu_l)(A_{min} - 1) + A_{min} + 1} \tag{13}$$

$$A_{min} > \frac{MAX_{C_l}(k_l, \mu_l) - 1}{MAX_{C_l}(k_l, \mu_l) + 1} \tag{14}$$

Finally, in order to unify all the weights in the network $\mathcal{N}$ for a program $\mathcal{P}$, given the definition of $MAX_{\mathcal{P}}(k_1, ..., k_q, \mu_1, ..., \mu_q)$, it is sufficient that equations 15 and 16 are satisfied by W and A_{min}, respectively.

$$W \geq \frac{2}{\beta} \cdot \frac{ln\,(1 + A_{min}) - ln\,(1 - A_{min})}{MAX_{\mathcal{P}}(k_1, ..., k_q, \mu_1, ..., \mu_q)(A_{min} - 1) + A_{min} + 1} \tag{15}$$

$$A_{min} > \frac{MAX_{\mathcal{P}}(k_1, ..., k_q, \mu_1, ..., \mu_q) - 1}{MAX_{\mathcal{P}}(k_1, ..., k_q, \mu_1, ..., \mu_q) + 1} \tag{16}$$

As a result, if equations 15 and 16 are satisfied by W and A_{min}, respectively, then $\mathcal{N}$ computes $T_{\mathcal{P}}$.□

Example 4 *Consider the program $\mathcal{P} = \{A \leftarrow B, C, notD;\ A \leftarrow E, F;\ B \leftarrow\}$. Converting fact $B \leftarrow$ to rule $B \leftarrow T$ and applying the* Translation Algorithm, *we obtain the neural network $\mathcal{N}$ of figure 3. Firstly, we calculate $MAX_{\mathcal{P}}(k_1, ..., k_n, \mu_1, ..., \mu_n) = 3$ (step 1), and $A_{min} > 0.5$ (step 2). Then, taking for example $A_{min} = 0.6$, we obtain $W \geq 6.931/\beta$ (step 3). Taking, alternatively, $A_{min} = 0.7$, we obtain $W \geq 4.336/\beta$. Therefore, assuming for instance $A_{min} = 0.7$ and $h(x)$ as the standard bipolar semi-linear activation function ($\beta = 1$), if $W = 4.5$ then $\mathcal{N}$ computes the operator $T_{\mathcal{P}}$ of $\mathcal{P}$*[4].

[4]Note that a sound translation from $\mathcal{P}$ to $\mathcal{N}$ do not require all the weights in $\mathcal{N}$ to have the same absolute value. We unify the weights ($|W|$) for the sake of simplicity of the translation algorithm and to stay in accordance with previous work.

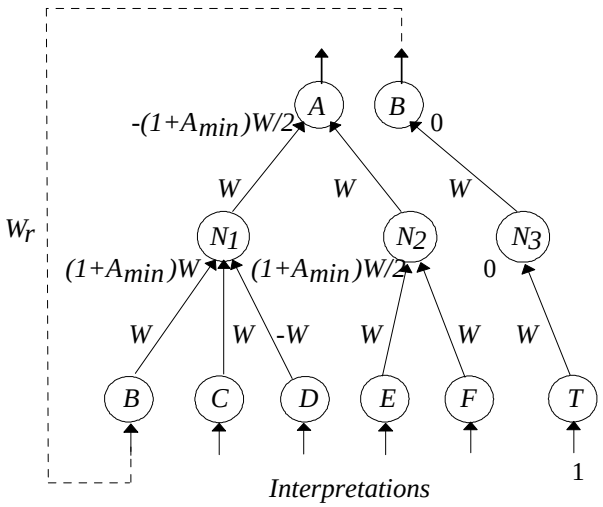

Figure 3: The neural network $\mathcal{N}$ obtained by the translation over $\mathcal{P}$. Connections with weight *zero* are not shown.

In the above example, the neuron B appears at both the input and the output layers of $\mathcal{N}$. That indicates that there are at least two clauses of $\mathcal{P}$ that are linked through B (in the example: $A \leftarrow B, C, notD$ and $B \leftarrow$), defining a *dependency chain* [3]. We represent that chain in the network using the recurrent connection $W_r = 1$ to denote that the output of B must feed the input of B in the next learning or recall step. In this way, regardless of the length of the dependency chains in $\mathcal{P}$, $\mathcal{N}$ always contains a single hidden layer and, consequently, a better learning performance can be obtained[5]. We will explain in detail the use of recurrent connections in section 3. We will compare the learning results of CIL^2P with *KBANN's*, where the number of hidden layers is equal to the length of the greatest dependency chain in the background knowledge, in section 4.

Remark 1 *Analogously to [23], for any logic program $\mathcal{P}$, the time needed to compute $T_{\mathcal{P}}(\mathbf{i})$ in the network is constant, equal to two time steps (one to compute the activations from the input to the hidden neurons and another from the hidden to the output neurons). A parallel computational model requiring $p(n)$ processors and $t(n)$ time to solve a problem of size n is optimal if $p(n) \times t(n) = O(T(n))$, where $T(n)$ is the best sequential time to solve the problem [26]. The number of neurons and connections in the network that corresponds to a program $\mathcal{P}$ is $O(q+r)$ and $O(q \cdot r)$, respectively, where q is the number of clauses and r is the number of propositional variables (atoms)*

[5] It is known that the increase in the number of hidden layers in a neural network creates a correspondent degradation in learning performance.

occurring in $\mathcal{P}$. The sequential time to compute $T_{\mathcal{P}}(\mathbf{i})$ is bound to $O(q \cdot r)$, and consequently, the above parallel computational model is optimal.

3 Massively Parallel Deduction and Inductive Learning

The neural network $\mathcal{N}$ can perform deduction and induction. In order to perform deduction, $\mathcal{N}$ is transformed into a partially recurrent network $\mathcal{N}_r$ by connecting each neuron in the output layer to its correspondent neuron in the input layer with weight $W_r = 1$, as shown in figure 4. In this way, $\mathcal{N}_r$ is used to iterate $T_{\mathcal{P}}$ in parallel, because its output vector becomes its input vector in the next computation of $T_{\mathcal{P}}$.

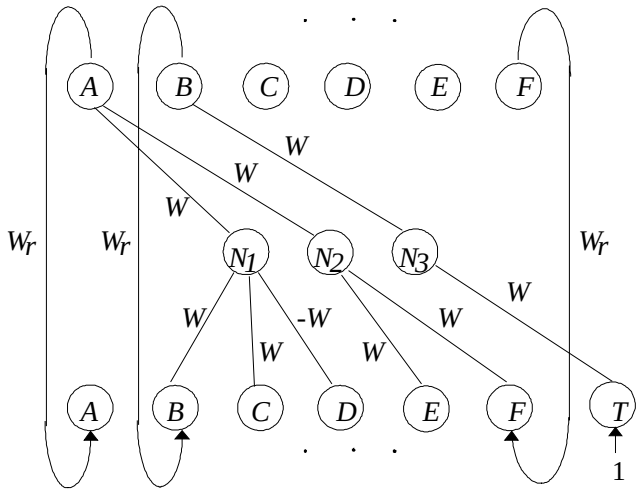

Figure 4: The recurrent neural network $\mathcal{N}_r$.

Let us now show that as in [23], if $\mathcal{P}$ is an acceptable program then $\mathcal{N}_r$ always settles down in a stable state that yields the unique fixed point of $T_{\mathcal{P}}$, since $\mathcal{N}_r$ computes the upward powers $(T^m_{\mathcal{P}}(\mathbf{i}))$ of $T_{\mathcal{P}}$. The same result could also be easily proved for the class of locally stratified programs (see [30]).

Definition 5 *[4, 3]: Let $B_{\mathcal{P}}$ denote the* Herbrand base *of $\mathcal{P}$, i.e. the set of propositional variables (atoms) occurring in $\mathcal{P}$. A* level mapping *for a program $\mathcal{P}$ is a function $|\ |: B_{\mathcal{P}} \rightarrow \aleph$ of ground atoms to natural numbers. For $A \in B_{\mathcal{P}}$, $|A|$ is called the* level *of A and $|notA| = |A|$.*

Definition 6 *[4, 3]: Let $\mathcal{P}$ be a program, $|\ |$ a level mapping for $\mathcal{P}$, and $\mathbf{i}$ a model of $\mathcal{P}$. $\mathcal{P}$ is called* acceptable *w. r. t $|\ |$ and $\mathbf{i}$ if for every clause*

$A \leftarrow L_1, ..., L_k$ *in* $\mathcal{P}$ *the following implication holds for* $1 \leq i \leq k$*: if* $\mathbf{i} \vDash \bigwedge_{j=1}^{i-1} L_j$ *then* $|A| > |L_j|$.

Theorem 7 *[14]: For each acceptable general program* $\mathcal{P}$*, the function* $T_\mathcal{P}$ *has a unique fixed-point. The sequence of all* $T_\mathcal{P}^m(\mathbf{i}), m \in \aleph$*, converges to this fixed-point* $T_\mathcal{P}^\varpi(\mathbf{i})$ *(which is identical to the stable model of* $\mathcal{P}$ *[18]), for each* $\mathbf{i} \subseteq B_\mathcal{P}$.

Remember that, since $\mathcal{N}_r$ has semi-linear neurons, for each real value o_i in the output vector ($\mathbf{o}$) of $\mathcal{N}_r$, if $o_i \geq A_{min}$ then the corresponding i-th atom in $\mathcal{P}$ is assigned "*true*", while $o_i \leq A_{max}$ means that it is assigned "$false$".

Corollary 8 *Let* $\mathcal{P}$ *be an acceptable general program. There exists a recurrent neural network* $\mathcal{N}_r$ *with semi-linear neurons such that, starting from an arbitrary initial input,* $\mathcal{N}_r$ *converges to a stable state and yields the unique fixed-point* $(T_\mathcal{P}^\varpi(\mathbf{i}))$ *of* $T_\mathcal{P}$*, which is identical to the stable model of* $\mathcal{P}$.
Proof. *Assume that* $\mathcal{P}$ *is an acceptable program. By theorem 3,* $\mathcal{N}_r$ *computes* $T_\mathcal{P}$*. Recurrently connected,* $\mathcal{N}_r$ *computes the upwards powers* $(T_\mathcal{P}^m(\mathbf{i}))$ *of* $T_\mathcal{P}$*. By theorem 7,* $\mathcal{N}_r$ *computes the unique stable model of* $\mathcal{P}$ $(T_\mathcal{P}^\varpi(\mathbf{i}))$.□

As a result, in order to use $\mathcal{N}$ as a massively parallel model for Logic Programming, we simply have to follow two steps: (i) add neurons to the input and output layers of $\mathcal{N}$, allowing it to be partially recurrently connected; and (ii) add the correspondent recurrent links with fixed weight $W_r = 1$.

Example 9 *(example 4 continued): Given any initial activation in the input layer of* $\mathcal{N}_r$ *(figure 4), it always converges to the following stable state:* $A =$ *"false",* $B =$ *"true",* $C =$ *"false",* $D =$ *"false",* $E =$ *"false", and* $F =$ *"false", that represents the unique stable model of* $\mathcal{P}$*,* $M(\mathcal{P}) = \{B\}$.

One of the main features of artificial neural networks is their learning capability. The program $\mathcal{P}$, viewed as background knowledge, may now be refined with examples in a neural training process on $\mathcal{N}$ or $\mathcal{N}_r$. Hornik et al. [24] have proved that standard feedforward neural networks with as few as a single hidden layer are capable of approximating any (Borel measurable) function from one finite dimensional space to another to any desired degree of accuracy, provided sufficiently many hidden units are available. Consequently, we can train single hidden layer neural networks to approximate the operator $T_\mathcal{P}$ associated with a logic program $\mathcal{P}$. Powerful neural learning algorithms have been established theoretically and applied extensively in practice. These algorithms may be used to learn the operator $T_{\mathcal{P}'}$ of a previously unknown program $\mathcal{P}'$ and, therefore, to learn the program $\mathcal{P}'$ itself. Moreover, DasGupta and Schinitger [11] have proved that neural networks with continuously differentiable activation functions are capable of

computing a certain family of boolean functions with constant size (n), while networks composed of binary threshold functions require at least $O(log(n))$ size. Hence, analog neural networks have more computational power than discrete neural networks, even when computing boolean functions.

The network's recurrent connections contain fixed weights $W_r = 1$, which exclusively represent the concept that the output should feed the input in the next learning or recall process. As $\mathcal{N}_r$ does not learn in its recurrent connections[6], the standard *backpropagation* learning algorithm can be applied directly [20] (see also [25]). Hence, in order to perform inductive learning with examples on $\mathcal{N}_r$, four simple steps should be followed: (i) add neurons to the input and output layers of $\mathcal{N}_r$, according to the training set (the training set may contain concepts not represented in the background knowledge and vice-versa); (ii) add neurons to the hidden layer of $\mathcal{N}_r$, if it is so required for the learning algorithm convergence; (iii) add connections with weight zero, in which $\mathcal{N}_r$ will learn new concepts; and (iv) perturb the connections by adding small random numbers to its weights in order to avoid learning problems caused by symmetry in $\mathcal{N}_r$[7]. The implementation of steps $(i) - (iv)$ will become clearer in section 4, where we describe some applications of the *C-IL^{2}P* system using *backpropagation*.

Remark 2 *The final stage of the* C-IL^{2}P *system is the symbolic knowledge extraction from the trained network. It is generally accepted that "*rules' extraction*" algorithms can provide the so called explanation capability for trained neural networks. The lack of explanation for their reasoning mechanisms is one of neural networks' main drawbacks. Similarly, the lack of clarity of trained networks has been the main reason for serious criticisms. The extraction of symbolic knowledge from trained networks can ameliorate considerably these problems. It makes the knowledge learned accessible for an expert's analysis and allows the justification of the decision making process. The knowledge extracted can be directly added to the knowledge base or used in the solution of analogous domains problems.*

Symbolic knowledge extraction from trained networks is an extensive topic by its own. Some of the main extraction proposals include [23, 2, 43, 54, 10, 16, 57, 50] (see [1] for a comprehensive survey). In the context of the C-IL^{2}P *system*[8]*, we have developed an extraction proposal [7] which main characteristic is that the extraction algorithm is* sound *and* complete[9].

[6] The recurrent connections represent an external process, between output and input only.

[7] The perturbation should be small enough not to have any effects on the computation of the background knowledge.

[8] In the *C-IL^{2}P* system, after learning $\mathcal{N}$ encodes a knowledge $\mathcal{P}'$ that contains the background knowledge $\mathcal{P}$ complemented or even revised by the knowledge learned with the training examples. $\mathcal{N}$ now computes $T_{\mathcal{P}'}$, instead of $T_{\mathcal{P}}$. Hence, an accurate extraction procedure must derive $\mathcal{P}'$ from $\mathcal{N}$.

[9] Briefly, soundness and completeness of an extraction algorithm are defined as follows

Moreover, the algorithm's search space can be reduced considerably (in the best case, the algorithm's complexity is linear) because we have identified a partial ordering in the input vectors set, over which a number of pruning rules can be applied safely. We have also defined a number of simplification rules that help reduce the length of the rule set, enhancing its readability and clarity. Since the algorithm is sound and complete, it guarantees that the extraction procedure approximates the network's exact representation and, in a particular application, the algorithm can be halted when a given degree of accuracy is obtained. The extraction step of C-IL^{2}P *is out of the scope of this paper, and the interested reader is referred to [7]. However, we would like to point out that there is a major conceptual difference between our approach and other extraction methods. We are convinced that an extraction method must consider* default negation *in the final rule set, and not only "*if then else*" rules. The network's behavior is nonmonotonic, and therefore we can not expect to map it properly into a set of rules composed of Horn clauses only.*

4 Experimental Results

We have applied the *C-IL^{2}P* system in two real-world problems in the domain of Molecular Biology, in particular DNA sequence analysis, the "*promoter recognition*" and the "*splice-junction determination*" problems[10]. Molecular Biology is an area of increasing interest for computational learning systems analysis and application. Specifically, DNA sequence analysis problems have recently become benchmark for learning systems' performance comparison. In this section, we will compare the experimental results obtained by *C-IL^{2}P* with a variety of learning systems.

In the sequel, we briefly introduce the problems in question from a computational application perspective (see [60] for a proper treatment on the subject). A DNA molecule contains two strands that are linear sequences of nucleotides. The DNA is composed from four different nucleotides - *adenine, guanine, thymine, and cytosine* - which are abbreviated by $\{a, g, t, c\}$, respectively. Some sequences of the DNA strand, called genes, serve as blueprint for the synthesis of proteins. Arranged between the genes are segments, called non-coding regions, that do not encode proteins.

Following [58], we use a special notation to identify the location of nucleotides in a DNA sequence. Each nucleotide is numbered with respect to a fixed, biologically meaningful, reference point. Rules' antecedents of

(see [16]): each rule r_i either belongs to the rule set or is subsumed by a rule that belongs to the rule set iff r_i is computed by the network.

[10] These are the same problems that were investigated in [58] for the evaluation of *KBANN*. We have followed as much as possible the methodology used by Towell and Shavlik, and we have used exactly the same background knowledge and set of examples as *KBANN*.

the form "*@3 atcg*" state the location relative to the reference point in the DNA, followed by the sequence of symbols that must occur. For example, "*@3 atcg*" means that an a must appear three nucleotides on the right of the reference point, followed by a t four nucleotides on the right of the reference point and so on. By convention, location zero is not used, while '$*$' means that any nucleotide will suffice in a particular location. In that way, a rule of the form $Minus35 \leftarrow @-36\text{'}ttg*ca\text{'}$ is a short representation for $Minus35 \leftarrow @-36\text{'}t\text{'}$, $@-35\text{'}t\text{'}$, $@-34\text{'}g\text{'}$, $@-32\text{'}c\text{'}$, $@-31\text{'}a\text{'}$. Each location is encoded in the network by four input neurons, representing nucleotides $\{a, g, t, c\}$ in that order. Rules are, therefore, inserted in the network as depicted in figure 5 for the hypothetical rule $Minus5 \leftarrow @-1\text{'}gc\text{'}$, $@5\text{'}t\text{'}$.

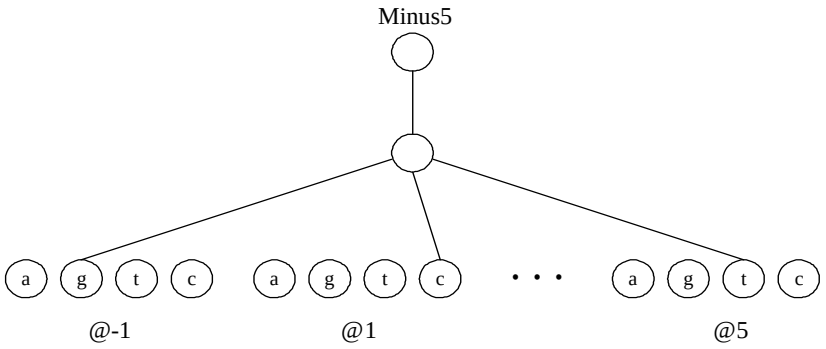

Figure 5: Inserting rule $Minus5 \leftarrow @-1\text{'}gc\text{'}$, $@5\text{'}t\text{'}$ into the neural network.

In addition to the reference point notation, Table 1 specifies a standard notation for referring to all possible combinations of nucleotides using a single letter. This notation is compatible with the EMBL, GenBank, and PIR data libraries, three major collections of data for molecular biology.

Code	Meaning	Code	Meaning	Code	Meaning
m	a *or* c	r	a *or* g	W	a *or* t
s	c *or* g	y	c *or* t	K	g *or* t
v	a *or* c *or* g	h	a *or* c *or* t	D	a *or* g *or* t
b	c *or* g *or* t	x	a *or* g *or* c *or* t		

Table 1: Single-letter codes for expressing uncertain DNA sequence.

The first application in which we test *C-IL*^{2}*P* is the prokaryotic[11] promoter recognition. Promoters are short DNA sequences that precede the beginning of genes. The aim of "*promoter recognition*" is to identify the

[11] Prokaryotes are single-celled organisms that do not have a nucleus, e.g. E. Coli.

starting location of genes in long sequences of DNA[12]. Table 2 contains the background knowledge for promoter recognition[13].

$Promoter \leftarrow Contact, Conformation$	
$Contact \leftarrow Minus10, Minus35$	
$Minus10 \leftarrow @-14$'$tataat$'	$Minus35 \leftarrow @-37$'$cttgac$'
$Minus10 \leftarrow @-13$'$tataat$'	$Minus35 \leftarrow @-36$'$ttgaca$'
$Minus10 \leftarrow @-13$'ta_*a_*t'	$Minus35 \leftarrow @-36$'$ttgac$'
$Minus10 \leftarrow @-12$'$ta_{***}t$'	$Minus35 \leftarrow @-36$'ttg_*ca'
$Conformation \leftarrow @-45$'$aa_{**}a$'	
$Conformation \leftarrow @-45$'$a_{***}a$', $@-28$'$t_{***}t_*aa_{**}t$', $@-4$'t'	
$Conformation \leftarrow @-49$'$a_{****}t$', $@-27$'$t_{****}a_{**}t_*tg$', $@-1$'a'	
$Conformation \leftarrow @-47$'$caa_*tt_*ac$', $@-22$'$g_{***}t_*c$', $@-8$'$gcgcc_*cc$'	

Table 2: Background knowledge for promoter recognition.

The background knowledge of Table 2 is translated by *C-IL^{2}P's* translation algorithm to the neural network of figure 6. In addition, two hidden neurons were added in order to facilitate the learning of new concepts from examples. Note that the network is fully-connected, but low-weighted links are not shown in the figure. The network's input vector for this task contains 57 consecutive DNA nucleotides. The training examples consist of 53 promoters and 53 nonpromoters DNA sequences.

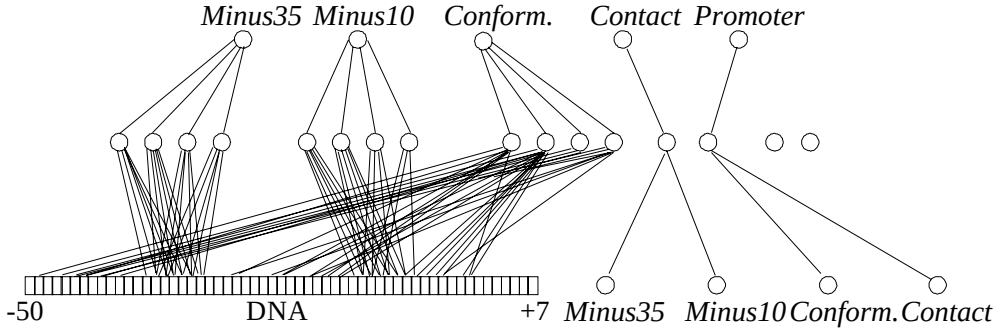

Figure 6: Initial neural network for promoter recognition. Each box at the input layer represents one sequence location that is encoded by four input neurons $\{a, g, t, c\}$.

[12] Promoters can be identified in laboratory by finding the location at which the RNA polymerase protein binds to the DNA sequence.

[13] Rules obtained from [58], and derived from the biological literature [38] from Noordewier [55].

The second application that we test *C-IL^{2}P* is eukaryotic[14] splice-junction determination. Splice-junctions are points on a DNA sequence at which the non-coding regions are removed during the process of protein synthesis. The aim of "*splice-junction determination*" is to recognize the boundaries between the part of the DNA retained after splice - called exons - and the part that is spliced out - the introns. The task consists, therefore, of recognizing exon/intron (E/I) boundaries and intron/exon (I/E) boundaries. Table 3 contains the background knowledge for splice junction determination[15].

$EI \leftarrow @-3$'$maggtragt$', $not\ EI - Stop$		
$EI - Stop \Leftarrow @-3$'taa'	$EI - Stop \Leftarrow @-4$'taa'	$EI - Stop \Leftarrow @-5$'taa'
$EI - Stop \Leftarrow @-3$'tag'	$EI - Stop \Leftarrow @-4$'tag'	$EI - Stop \Leftarrow @-5$'tag'
$EI - Stop \Leftarrow @-3$'tga'	$EI - Stop \Leftarrow @-4$'tga'	$EI - Stop \Leftarrow @-5$'tga'

$IE \leftarrow pyramidine - rich, @-3$'$yagg$', $not\ IE - Stop$		
$pyramidine - rich \leftarrow 6\ of\ (@-15$'$yyyyyyyyyy$')		
$IE - Stop \Leftarrow @1$'taa'	$IE - Stop \Leftarrow @2$'taa'	$IE - Stop \Leftarrow @3$'taa'
$IE - Stop \Leftarrow @1$'tag'	$IE - Stop \Leftarrow @2$'tag'	$IE - Stop \Leftarrow @3$'tag'
$IE - Stop \Leftarrow @1$'tga'	$IE - Stop \Leftarrow @2$'tga'	$IE - Stop \Leftarrow @3$'tga'

Table 3: Background knowledge for splice-junction.

The background knowledge of Table 3 is translated by *CIL^{2}P* to the neural network of figure 7. In Table 3, "$\Leftarrow$" indicates nondefeasible rules, which can not be altered during training. Therefore, the weights set in the network by these rules are fixed. Rules of the form "$m\ of\ (...)$" are satisfied if at least m of the parenthesized concepts are true. Note that the translation of these rules to the network is done by simply defining $k_l = m$ in *CIL^{2}P*'s translation algorithm. Rules containing symbols other than the original $\{a, g, t, c\}$ are split into a number of equivalent rules containing only the original symbols, according to Table 1. For instance, $IE \leftarrow pyramidine - rich, @-3$'$yagg$', $not\ IE - Stop$ is encoded in the network as $IE \leftarrow pyramidine - rich, @-3$'$cagg$', $not\ IE - Stop$ and $IE \leftarrow pyramidine - rich, @-3$'$tagg$', $not\ IE - Stop$, since $y \equiv c \vee t$.

The training set for this task contains 3190 examples, in which approximately 25% are of I/E boundaries, 25% are of E/I boundaries and the remaining 50% are neither. The third category (neither E/I nor I/E) is considered true when neither I/E nor E/I output neurons are active. Each example is a DNA sequence with 60 nucleotides, where the center is the reference point. Remember that the network of figure 7 is fully-connected,

[14] Unlike prokaryotic cells, eukaryotic cells have a nucleus, what characterize more developed organisms.

[15] Rules obtained from [58] and derived from the biological literature from Noordewier [37].

but that low-weighted links are not shown. Doted lines indicate links with negative weights.

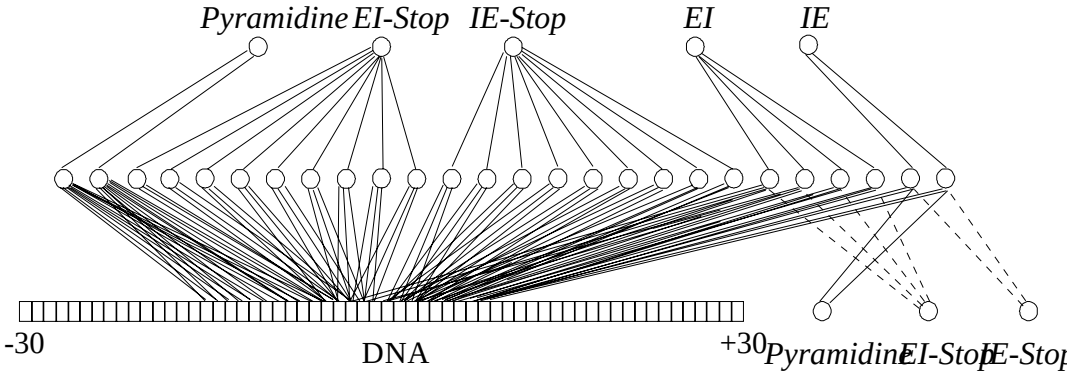

Figure 7: Initial neural network for splice-junction determination. Each box at the input layer of the network represents one sequence location which is encoded by four input neurons $\{a, g, t, c\}$.

Let us now describe the experimental results obtained by *C-IL²P* in the applications above. We compare it with other symbolic, neural and hybrid learning systems. Briefly, our tests show that *C-IL²P* is a very effective system. Its test set performance is at least as good as *KBANN's* and, therefore, better than any method analyzed in [58]. Moreover, *C-IL²P*'s training set performance is considerably superior than *KBANN's*, mainly because it always encodes the background knowledge in a single hidden layer network.

Firstly, let us consider *C-IL²P*'s test-set performance, that is, its ability to generalize over examples not seen during training. We compare the results obtained by *C-IL²P* in both applications with some of the main inductive learning systems from examples: Backpropagation [48], Perceptron [47] (neural systems), ID3 [44], and Cobweb [13] (symbolic systems). We also compare the results in the promoter recognition problem with a method suggested by biologists [51]. In addition, we compare *C-IL²P* with systems that learn from both examples and background knowledge: Either [39], Labyrinth-K [52], FOCL [40] (symbolic systems), and KBANN [58] (hybrid system)[16].

As in [58], we evaluate the systems using *cross-validation*, a testing methodology in which the set of examples is permuted and divided into n sets. One division is used for testing and the remaining $n - 1$ divisions

[16] Towell and Shavlik compare *KBANN* with other hybrid systems [15] and [27], obtaining better results.

are used for training. The testing division is never seen by the learning algorithm during the training process. The procedure is repeated n times so that every partition is used once for testing. For the 106-examples promoter data set, we use leaving-one-out cross-validation, in which each example is successively left out of the training set. Hence, it requires 106 training phases, in which the training set has 105 examples and the testing set has 1 example. Leaving-one-out becomes expensive as the number of available examples grows. Therefore, following [58], we use 10-fold cross-validation for the 1000-examples splice-junction determination data set[17].

The learning systems that are based on neural networks have been trained until one of the following three stopping criteria was satisfied: (i) on 99% of the training examples, the activation of every output unit is within 0.25 of correctness; (ii) every training example is presented to the network 100 times, that is, the network has been trained for 100 epochs; or (iii) the network classifies at least 90% of the training examples correctly but has not improved its classification ability for 5 epochs. We have defined an epoch as one training pass through the whole training set. We used the standard backpropagation learning algorithm to train *C-IL*^{2}*P*'s networks.

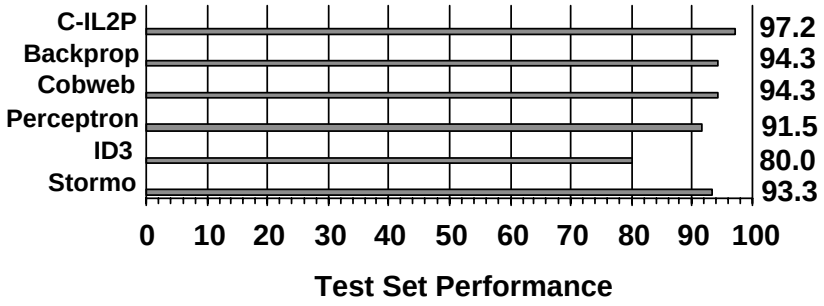

Figure 8: Test-set performance in the promoter problem (comparison with systems that learn strictly from examples).

C-IL^2*P* generalizes better than any empirical learning system (see figures 8 and 9) and than any system that learns from examples and background knowledge (see figures 10 and 11) tested on both applications. In most cases, differences are statistically significant. However, *C-IL*^{2}*P* is only marginally better than *KBANN*. That results from the fact that both systems are hybrid neural systems that perform inductive learning from examples and background knowledge.

[17] In accordance with [58], 1000 examples are randomly selected from the 3190 examples set for each training phase.

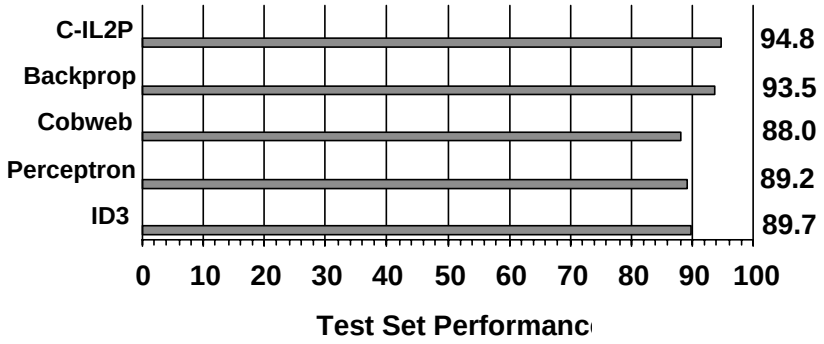

Figure 9: Test-set performance in the splice junction problem (comparison with systems that learn strictly from examples).

Usually, theory and data learning systems require fewer training examples than systems that learn only from data. The background knowledge helps a learning system to extract useful generalizations from small sets of examples. That is quite important since, in general, it is not easy to obtain large and accurate training sets.

Let us now analyze *C-IL^{2}P*'s test-set performance given smaller sets of examples. The following tests will compare the performance of *C-IL^{2}P* with *KBANN* and *Backpropagation* only, because these systems have shown to be the most effective ones in the previous tests (figures 8, 9, 10 and 11). Following [58], the generalization ability over small sets of examples is analyzed by splitting the examples into two subsets: the testing set containing approximately 25% of the examples, and the training set containing the remaining examples. The training set is partitioned into sets of increasing sizes and the networks are trained using each partition at a time.

Figures 12 and 13 show that, in both applications, *C-IL^{2}P* generalizes over small sets of examples better than *backpropagation*. The results empirically show that the initial topology of the network, set by the background knowledge, confers it a better generalization capability. Note that the results obtained by *C-IL^{2}P* and *KBANN* are very similar, since both systems are based on the backpropagation learning algorithm and learn from examples and background knowledge.

Concluding the tests, we check the training-set performance of *C-IL^{2}P* in comparison again with *KBANN* and *backpropagation*. Figures 14 and 15 describe the training-set RMS error rate decay obtained by each system during learning, respectively in each application. The RMS parameter represents how fast a neural network learns a set of examples w.r.t training epochs. Neural networks learning performance is a major concern, since it

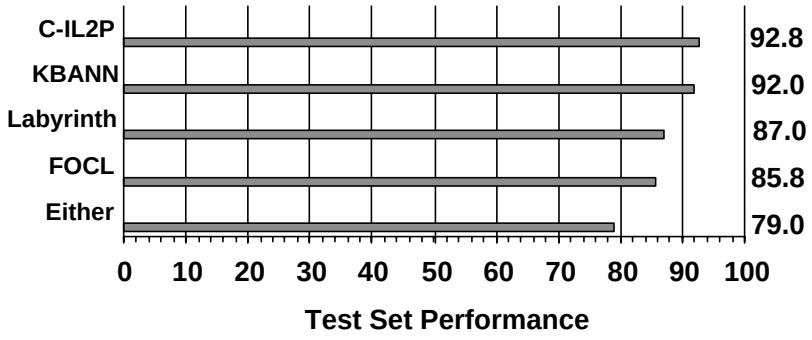

Figure 10: Test-set performance in the promoter problem (comparison with systems that learn both from examples and theory).

can become prohibitive in certain applications, usually as a result of the local minima problem[18].

Figures 14 and 15 show that *C-IL^{2}P's* learning performance is considerably better than *KBANN's*. The results suggest that our translation algorithm from symbolic knowledge to neural networks is more adequate than the one presented in [58]. The *Translation Algorithm* presented here always encodes the background knowledge into a single hidden-layer neural network. Conversely, *KBANN's* translation algorithm generates a network with as many layers as dependency chains there are in the background knowledge. For example, if $B \leftarrow A, C \leftarrow B, D \leftarrow C, and\ E \leftarrow D$, *KBANN* generates a network with three hidden-layers in which concepts B, C, and D are represented. Obviously, that creates a respective degradation in learning performance. Towell and Shavlik have tried to overcome this problem with a symbolic pre-processor of rules for *KBANN* [59]. However, it creates another preliminary phase to their translation process[19]. In our point of view, the problem lies in *KBANN's* translation algorithm, and can be straightforwardly solved by an accurate translation mechanism.

Summarizing, the experiments described above suggest that *C-IL^{2}P*'s effectiveness is given by three system's features: *CIL^{2}P* is based on *backpropagation*, it uses *background knowledge*, and it provides an *accurate and compact translation* from symbolic knowledge to neural networks.

[18] The network can get stuck in local minima during learning, instead of finding the global minimum of its error function.

[19] *KBANN* already contains a preliminary phase of rules hierarchying, that rewrites the rules before translating them.

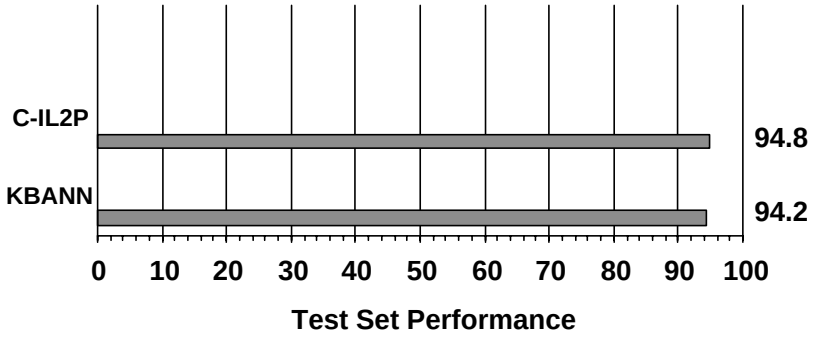

Figure 11: Test-set performance in the splice junction problem (comparison with systems that learn both from examples and theory).

5 Future Work and Conclusion

There are some interesting open questions related to the *explanation capability* of neural networks, specially to the trade-off between *complexity* and *quality* of rules' extraction methods. An alternative to reduce this trade-off is to investigate more efficient pruning methods for neural networks' input vectors search space.

Another interesting question is related to the class of extended programs [19], that is of interest in connection with the relation between Logic Programming and nonmonotonic formalisms. *Extended logic programs*, that add "classical" negation to the language of general programs, can be viewed as a fragment of Default theories [46]. Some facts of commonsense knowledge can be represented more easily when "classical" negation is available. We have extended *C-IL*^{2}*P* to deal with extended logic programs. The extended *C-IL*^{2}*P* system computes *answer sets* [19], instead of stable models. As a result, it can be applied in a broader range of domains theories. The extended *C-IL*^{2}*P* was already applied in power systems' fault diagnosis, obtaining promising preliminary results [6].

As a massively parallel nonmonotonic learning system, *C-IL*^{2}*P* is of interest in relation with *Belief Revision*. In one of the above applications, the background knowledge was revised by the set of examples, because it was defeasible and some examples contradicted it. Therefore, background knowledge and set of examples can be inconsistent, and one needs to investigate ways to detect and treat *inconsistencies* in the system, looking at the learning process as a revision one.

By changing the definition of $T_{\mathcal{P}}$, variants of Default Logic and Logic Pro-

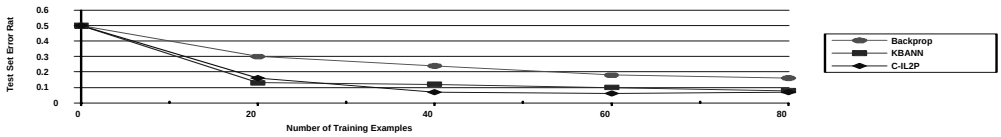

Figure 12: Test-set error rate in the promoter problem (26 examples reserved for testing).

gramming semantics can be obtained [61], defining a family of *nonmonotonic neural reasoning systems*. Another interesting question is to investigate if by using *labelled clauses*, similarly to [8] and more generally to [17], one can obtain the proof of a literal as its label. The network's learnability and generalizability must also be formally studied, having Logic as the underlying formalism. The system's extension to deal with *first order logic* is another complex and vast area for further research.

In this paper, we have presented the Connectionist Inductive Learning and Logic Programming System (*C-IL^{2}P*), that is a massively parallel computational model based on artificial neural networks that integrates inductive learning from examples and background knowledge with deductive learning from Logic Programming. We have obtained successful experimental results applying *C-IL^{2}P* in two real-world problems in the molecular biology domain. "*Both kinds of Intelligent Computational Systems, Symbolic and Connectionist, have virtues and deficiencies. It is very important to integrate them, through neural-symbolic systems, in order to explore the capabilities each one possesses*" [34]. In this sense, the integration that we have presented here is tightly coupled [32, 21], expecting to have contributed to this long term research.

Acknowledgements

We are grateful to Valmir Barbosa, Rodrigo Basilio, Dov Gabbay, Luis Alfredo Carvalho and Luis Lamb for useful discussions. We would especially like to thank Krysia Broda and Sanjay Modgil for their comments. This research was partially supported by CNPq and CAPES/Brazil.

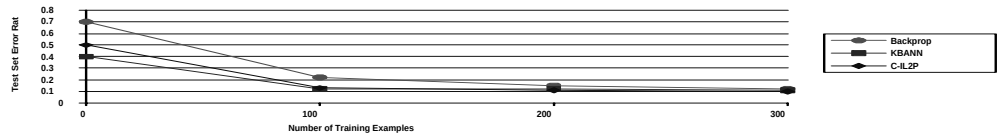

Figure 13: Test-set error rate in the splice junction problem (798 examples reserved for testing).

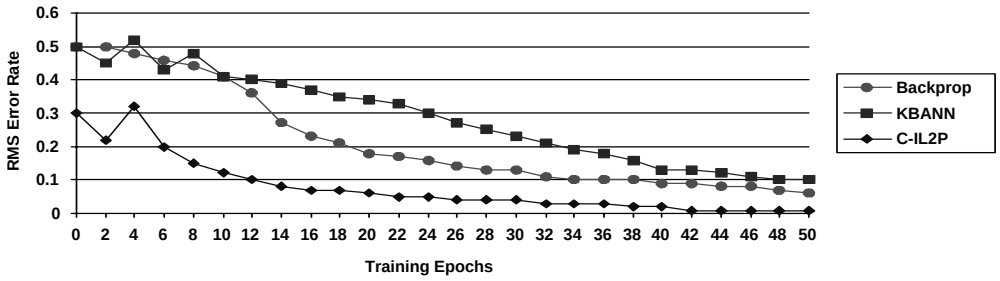

Figure 14: Training-set RMS error decay during learning the promoter problem.

References

[1] R. Andrews, J. Diederich and A. B. Tickle, *A Survey and Critique of Techniques for Extracting Rules from Trained Artificial Neural Networks*, Knowledge-based Systems, 8(6):1-37, 1995.

[2] R. Andrews and S. Geva, *Inserting and Extracting Knowledge from Constrained Error Backpropagation Networks*, In: Proc. 6^{th} Australian Conference on Neural Networks, Sydney, 1995.

[3] K. R. Apt and N. Bol, *Logic Programming and Negation: A Survey*, Journal of Logic Programming, 19:9-71, 1994.

[4] K. R. Apt and D. Pedreschi, *Reasoning about Termination of Pure Prolog Programs*, Information and Computation, 106:109-157, 1993.

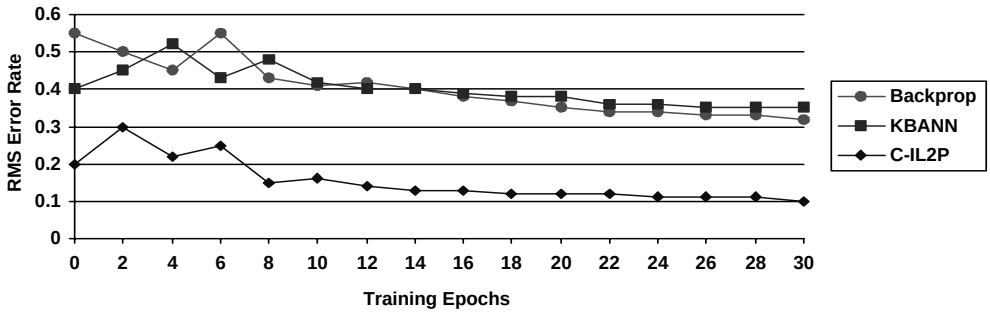

Figure 15: Training-set RMS error decay during learning the splice junction problem.

[5] A. d'Avila Garcez, G. Zaverucha and L. A. Carvalho, *Logic Programming and Inductive Learning in Artificial Neural Networks*, In: Ch. Herrmann, F. Reine and A. Strohmaier (eds.), Knowledge Representation in Neural Networks, Logos-Verlag Berlin, 33-46, 1997.

[6] A. d'Avila Garcez, G. Zaverucha and V. da Silva, *Applying the Connectionist Inductive Learning and Logic Programming System to Power System Diagnosis*, In: Proc. IEEE International Joint Conference on Neural Networks ICNN'97, 1:121-126, Houston, 1997.

[7] A. d'Avila Garcez, K. Broda and D. Gabbay, *Symbolic Knowledge Extraction from Trained Neural Networks*, Technical Report, Department of Computing, Imperial College, London, 1998.

[8] G. Brewka, *Cumulative Default Logic - In defense of nonmonotonic inference rules*, Artificial Intelligence, 50(2):183-205, 1991.

[9] N. K. Bose and P. Liang, *Neural Networks Fundamentals with Graphs, Algorithms, and Applications*, McGraw-Hill, 1996.

[10] M. W. Craven and J. W. Shavlik, *Using Sampling and Queries to Extract Rules from Trained Neural Networks*, in: Proc. Eleventh International Conference on Machine Learning, 37-45, 1994.

[11] B. DasGupta and G. Schinitger, *Analog Versus Discrete Neural Networks*, Neural Computation, 8:805-818, 1996.

[12] H. B. Enderton, *A Mathematical Introduction to Logic*, Academic Press, 1972.

[13] D. H. Fisher, *Knowledge Acquisition via Incremental Conceptual Clustering*, Machine Learning, 2:139-172, 1987.

[14] M. Fitting, *Metric Methods - Three Examples and a Theorem*, Journal of Logic Programming, 21:113-127, 1994.

[15] L. M. Fu, *Integration of Neural Heuristics into Knowledge-based Inference*, Connection Science, 1:325-340, 1989.

[16] L. M. Fu, *Neural Networks in Computer Intelligence*, McGraw Hill, 1994.

[17] D. M. Gabbay, *LDS - Labelled Deductive Systems - Volume 1 Foundations*, Oxford University Press, 1996.

[18] M. Gelfond and V. Lifschitz, *The Stable Model Semantics for Logic Programming*, In: Proc. 5th International Symposium on Logic Programming, MIT Press, 1070-1080, Cambridge, 1988.

[19] M. Gelfond and V. Lifschitz, *Classical Negation in Logic Programs and Disjunctive Databases*, New Generation Computing, 9:365-385, Springer-Verlag, 1991.

[20] J. Hertz, A. Krogh and R. G. Palmer, *Introduction to the Theory of Neural Computation*, Santa Fe Institute, Studies in the Science of Complexity, Addison-Wesley Publishing Company, 1991.

[21] M. Hilario, *An Overview of Strategies for Neurosymbolic Integration*, In: Proc. Workshop on Connectionist-Symbolic Integration: from Unified to Hybrid Approaches, IJCAI 95, 1995.

[22] S. Holldobler, *Automated Inferencing and Connectionist Models*, Postdoctoral Thesis, Intellektik, Informatik, TH Darmstadt, 1993.

[23] S. Holldobler and Y. Kalinke, *Toward a New Massively Parallel Computational Model for Logic Programming*, in: Proc. Workshop on Combining Symbolic and Connectionist Processing, ECAI 94, 1994.

[24] K. Hornik, M. Stinchcombe and H. White, *Multilayer Feedforward Networks are Universal Approximators*, Neural Networks, 2:359-366, 1989.

[25] M. I. Jordan, *Attractor Dynamics and Parallelisms in a Connectionist Sequential Machine*, In: Proc. Eighth Annual Conference of the Cognitive Science Society, 531-546, 1986.

[26] R. M. Karp and V. Ramachandran, *Parallel Algorithms for Shared-Memory Machines*, In: J. van Leeuwen (ed.), Handbook of Theoretical Computer Science, 17:869-941, Elsevier Science, 1990.

[27] B. F. Katz, *EBL and SBL: A Neural Network Synthesis*, In: Proc. Eleventh Annual Conference of the Cognitive Science Society, 683-689, Ann Arbor, 1989.

[28] F. J. Kurfeß, *Neural Networks and Structured Knowledge*, In: Ch. Herrmann, F. Reine and A. Strohmaier (eds.), Knowledge Representation in Neural Networks, Logos-Verlag Berlin, 5-22, 1997.

[29] N. Lavrac and S. Dzeroski, *Inductive Logic Programming: Techniques and Applications*, Ellis Horwood Series in Artificial Intelligence, 1994.

[30] J. W. Lloyd, *Foundations of Logic Programming*, Springer - Verlag, 1987.

[31] W. Marek and M. Truszczynski, *Nonmonotonic Logic: Context Dependent Reasoning*, Springer-Verlag, 1993.

[32] L. Medsker, *Neural Networks Connections to Expert Systems*, In: Proc. World Congress on Neural Networks, 411-417, 1994.

[33] R. S. Michalski, *Learning Strategies and Automated Knowledge Acquisition*, Computational Models of Learning, Symbolic Computation, Springer-Verlag, 1987.

[34] M. Minsky, *Logical versus Analogical, Symbolic versus Connectionist, Neat versus Scruffy*, AI Magazine, 12(2), 1991.

[35] T. M. Mitchell, *Machine Learning*, McGraw-Hill, 1997.

[36] S. Muggleton and L. Raedt, *Inductive Logic Programming: Theory and Methods*, Journal of Logic Programming, 19:629-679, 1994.

[37] M. O. Noordewier, G. G. Towell and J. W. Shavlik, *Training Knowledge-based Neural Networks to recognize genes in DNA sequences*, In: Advances in Neural Information Processing Systems, 3:530-536, Denver, 1991.

[38] M. C. O'Neill, *Escherichia Coli Promoters: Consensus as it relates to spacing class, cpecificity, repeat substructure, and three dimensional organization*, Journal of Biological Chemistry, 264:5522-5530, 1989.

[39] D. Ourston and R. J. Mooney, *Theory Refinement Combining Analytical and Empirical Methods*, Artificial Intelligence, 66:273-310, 1994.

[40] M. Pazzani and D. Kibler, *The Utility of Knowledge in Inductive Learning*, Machine Learning, 9:57-94, 1992.

[41] G. Pinkas, *Energy Minimization and the Satisfiability of Propositional Calculus*, Neural Computation, 3(2), 1991.

[42] G. Pinkas, *Reasoning, Nonmonotonicity and Learning in Connectionist Networks that Capture Propositional Knowledge*, Artificial Intelligence, 77:203-247, 1995.

[43] E. Pop, R. Hayward and J. Diederich, *RULENEG: Extracting Rules from a Trained ANN by Stepwise Negation*, QUT NRC, 1994.

[44] J. R. Quinlan, *Induction of Decision Trees*, Machine Learning, 1: 81-106, 1986.

[45] J. R. Quinlan, *Learning Logical Definitions from Relations*, Machine Learning, 5:239-266, 1990.

[46] R. Reiter, *A Logic for Default Reasoning*, Artificial Intelligence, 13:81-132, 1980.

[47] F. Rosenblatt, *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*, Spartan Books, New York, 1962.

[48] D. E. Rumelhart, G. E. Hinton and R. J. Williams, *Learning Internal Representations by Error Propagation*, In: D. E. Rumelhart and J. L. McClelland (eds.), Parallel Distributed Processing, 1:318-363, MIT Press, 1986.

[49] R. Setiono, *A Penalty-function for Pruning Feedforward Neural Networks*, Neural Computation, 9:185-204, 1997.

[50] R. Setiono, *Extracting Rules from Neural Networks by Pruning and Hidden-unit Splitting*, Neural Computation, 9:205-225, 1997.

[51] G. D. Stormo, *Consensus Patterns in DNA*, Methods in Enzymology, 183:211-221, Academic Press, Orlando, 1990.

[52] K. Thompson, P. Langley and W. Iba, *Using Background Knowledge in Concept Formation*, In: Proc. Eighth International Machine Learning Workshop, 554-558, Evanston, 1991.

[53] S. B. Thrun, J. Bala, E. Bloedorn, I. Bratko, B. Cestnik, J. Cheng, K. De Jong, S. Dzeroski, S. E. Fahlman, D. Fisher, R. Haumann, K. Kaufman, S. Keller, I. Kononenko, J. Kreuziger, R. S. Michalski, T. Mitchell, P. Pachowicz, Y. Reich, H. Vafaie, K. Van de Welde, W. Wenzel, J. Wnek and J. Zhang, *The MONK's Problem: A Performance Comparison of Different Learning Algorithms*, Technical Report, Carnegie Mellon University, 1991.

[54] S. B. Thrun, *Extracting Provably Correct Rules from Artificial Neural Networks*, Technical Report, Institut fur Informatik, Universitat Bonn, 1994.

[55] G. G. Towell, J. W. Shavlik and M. O. Noordewier, *Refinement of Approximately Correct Domain Theories by Knowledge-based Neural Networks*, In: Proc. AAAI'90, 861-866, Boston, 1990.

[56] G. G. Towell, *Symbolic Knowledge and Neural Networks: Insertion, Refinement and Extraction*, PhD Thesis, Computer Sciences Department, University of Wisconsin, Madison, 1991.

[57] G. G. Towell and J. W. Shavlik, *The Extraction of Refined Rules From Knowledge Based Neural Networks*, Machine Learning, 13(1):71-101, 1993.

[58] G. G. Towell and J. W. Shavlik, *Knowledge-Based Artificial Neural Networks*, Artificial Intelligence, 70(1):119-165, 1994.

[59] G. G. Towell and J. W. Shavlik, *Using Symbolic Learning to Improve Knowledge-Based Neural Networks*, In: Proc. AAAI'94, 1994.

[60] J. D. Watson, N. H. Hopkins, J. W. Roberts, J. A. Steitz and A. M. Weiner, *Molecular Biology of the Gene, Volume 1*, Benjamin Cummings, Menlo Park, 1987.

[61] G. Zaverucha, *A Prioritized Contextual Default Logic: Curing Anomalous Extensions with a Simple Abnormality Default Theory*, In: Proc. KI'94, LNAI 861, Springer-Verlag, 1994.