ELSEVIER

Artificial Intelligence in Medicine 28 (2003) 121–140

Artificial Intelligence in Medicine

www.elsevier.com/locate/artmed

Inductive bias strength in knowledge-based neural networks: application to magnetic resonance spectroscopy of breast tissues

Christian W. Omlin[a,*], Sean Snyders[b,1]

[a]*Department of Computer Science, University of the Western Cape, 7535 Bellville, South Africa*
[b]*Department of Computer Science, University of Stellenbosch, 7600 Stellenbosch, South Africa*

Received 30 April 2001; received in revised form 14 April 2003; accepted 6 May 2003

Abstract

The integration of symbolic knowledge with artificial neural networks is becoming an increasingly popular paradigm for solving real-world applications. The paradigm provides means for using prior knowledge to determine the network architecture, to program a subset of weights to induce a learning bias which guide network training, and to extract knowledge from trained networks. The role of neural networks then becomes that of knowledge refinement. It thus provides a methodology for dealing with uncertainty in the prior knowledge. We address the open question of how to determine the strength of the inductive bias of programmed weights; we present a quantitative solution which takes the network architecture, the prior knowledge, and the training data into consideration. We apply our solution to the difficult problem of analyzing breast tissue from magnetic resonance spectroscopy (MRS); the available database is extremely limited and cannot be adequately explained by expert knowledge alone.
© 2003 Elsevier B.V. All rights reserved.

Keywords: Knowledge-based neural networks; Inductive bias; Learning with hints; Training and generalization performance; ^{31}P magnetic resonance spectroscopy; Breast tissue

1. Introduction

1.1. Expert knowledge and inductive learning

Medical decision making is well-suited for the application of artificial intelligence techniques [23,37]. Expert knowledge about disease processes is available which can be

* Corresponding author.
E-mail addresses: comlin@uwc.ac.za (C.W. Omlin), sean.snyders@i-u.de (S. Snyders).
[1] Present address: Computational Intelligence Group, School of Information Technology, International University, 76646 Bruchsal, Germany.

0933-3657/$ – see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0933-3657(03)00062-9

expressed in the form of rules. Such rules can be implemented in expert systems or used to guide analytical learning methods such as explanation-based learning [21]. The objective of analytical learning methods is to find a hypothesis which fits both the expert knowledge and the given examples of case histories. They provide logically justified hypotheses which have been arrived at through deductive inference. They have the advantage that they can learn from *sparse data*; however, the logical justifications are only as valid as the expert knowledge they are based on. They are of limited use of the expert knowledge is either incomplete or even incorrect.

Expert knowledge in the medical field is often incomplete due to the variability and the complexity of disease processes. Inductive systems such as decision trees and neural networks which are able to learn from case histories have been successfully applied in medicine [23]. They seek general hypotheses that fit the observed training data; thus, they rely on the statistical argument that the training sample is sufficiently large and thus representative of the distribution underlying the data. Theoretical results assert that large amounts of sample data are usually required for satisfactory performance of the trained models [17]. Acquisition of large amounts of data is often infeasible which leaves the trained models suspect.

Analytical and inductive learning work well under different conditions. However, most practical learning problems lie somewhere between the two extremes of plentiful data without prior knowledge and perfect prior knowledge with scarce data. Combining inductive with analytical learning methods thus holds the promise of exploiting the strengths of the two approaches while alleviating their respective weaknesses. This hybrid approach is applicable to many practical problems including computer-assisted medical diagnosis.

1.2. Learning bias in inductive methods

The choice of an appropriate machine learning method for modeling a problem is difficult; a wrong choice can make the learning process very difficult instead of increasing the learning performance or the predictive accuracy of the algorithm. Too many parameters in model-free inference (e.g. a neural network) can lead to high *variance* in the estimation error whereas model-based inference can introduce a bias which can hinder the search for a solution in hypothesis space. Model-based inference generally provides an inductive bias that restricts the search for a solution. However, a less restrictive model may give insufficient guidance towards a solution.

We can distinguish two types of inductive bias [9]: *representational bias* defines the states in the hypothesis search space; it can be introduced through some language (e.g. propositional logic) or structure (e.g. decision trees, neural network architectures); *procedural bias* defines the manner in which the hypothesis space is searched (e.g. high information gain attributes close to the root in decision trees, gradient-descent search in weight space of neural networks).

All machine learning methods have some inherent bias toward finding a solution in hypothesis space. For instance, the inductive bias of the ID3 algorithm for building decision trees is toward shallow trees that place high information gain attributes close to the root; the error backpropagation algorithm for feedforward neural networks is biased toward

finding a smooth interpolation between data points. However, these implicit biases are often not sufficient and an explicit bias must be introduced to achieve acceptable training and generalization performance.

Methods that adapt their biases have been proposed [43]. Hybrid methods that combine different learning paradigms have been developed [39]; generally, the combined methods have achieved results that are equal or better than the performance of the best of the individual methods [6]. The stability of algorithms and the effects of inductive biases on this stability have been discussed in [49]. Evaluation and selection of different biases has been investigated in the literature [9,34,50]. For an example of an implementation of a practical system, see [18].

1.3. The variance/bias dilemma for neural networks

It has been stated a long time ago that neural networks cannot be expected to learn anything useful without some significant prior structure [27]. Recent theoretical results support that point of view [17]. The impact of training feedforward neural networks with prior knowledge on the computational learning complexity has been discussed in [1]. The sample complexity for valid generalization has been investigated in [12]; the result shows that knowledge-based neural networks require a smaller sample size for valid generalization compared to networks trained without prior knowledge.

Learning with prior knowledge (also known as learning with hints) has attracted increasing attention. The philosophy of learning with hints is that since training neural networks is an inherently difficult problem, any and all prior knowledge that is available should be taken advantage of. One approach is to prestructure or initialize a network with knowledge prior to training [29,46]. The goal is to reduce training time and to improve network generalization performance. Thus, the role of neural networks then becomes that of *knowledge refinement* or even *knowledge revision* in the case where the prior knowledge is incorrect [31].

1.4. Overview

Following this general discussion about inductive bias in machine learning, in Section 2 we will focus our attention to knowledge-based neural networks. We will present a general framework for combining symbolic and neural learning, and present examples of medical applications where symbolic knowledge has been integrated with neural networks. We discuss knowledge-based artificial neural network (KBANN) as a representive example of such an integration in Section 3; KBANN maps prior knowledge in the form of propositional rules into feedforward networks thus providing an explicit inductive bias. We conclude that discussion with the open question: how strong should that inductive bias be chosen in order to achieve good training and generalization performance? We propose a heuristic for determining the strength of this inductive bias which takes into account the neural network architecture, the prior knowledge, and the training data in Section 4. We apply our heuristic to the problem of magnetic resonance spectroscopy of breast tissue in Section 5. We find that our heuristic for determining the strength of the inductive bias outperforms both average and standard choices for the

inductive bias. We end this paper with conclusions from our work and possible directions for future research.

2. Integrating symbolic knowledge with neural networks

2.1. General paradigm

Combining symbolic and neural learning has become a well-established paradigm [1,5,10,13,14,19,24–26,29,33,39,44,45,47,48]. There are different ways in which neural and symbolic learning can be combined to solve a given learning task. An excellent collection of a variety of approaches can be found in [11].

Following the discussion in Section 1.3, the question arises whether neural networks can make effective use of explicit inductive bias and how such a bias influences the training and generalization performance. In order to introduce an explicit bias in feedforward neural networks, i.e. a preference for a solution in weight space, one has to investigate how neural networks represent knowledge and infer hypotheses from learning examples. Weighted connections between neurons provide an opportunity to incorporate knowledge prior to learning. The structure of the network and the programmed weights provide an explicit inductive bias.

In the following, we will limit our discussion to the paradigm illustrated in Fig. 1. The traditional approach to using neural networks is shown in the lower part ('connectionist representation'). A network's adaptable weights are initialized with random values drawn

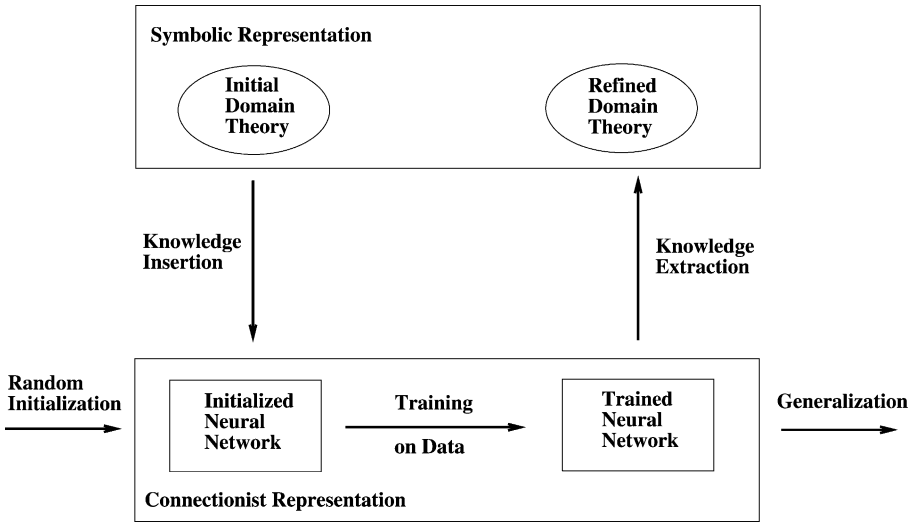


Fig. 1. A framework for combining symbolic and neural learning: the use of a neural network for knowledge refinement consists of three steps: (1) insertion of prior symbolic knowledge (initial domain theory) into a neural network, (2) refinement of knowledge through training a neural network on examples, and (3) extraction of symbolic or learned knowledge (refined domain theory) from a trained network.

according to some distribution. Using numerical optimization methods (e.g. gradient descent techniques, simulated annealing), the network is trained on some known data to perform a certain task (e.g. pattern classification) until some training criterion is met. After successful training, a network can take advantage of its generalization capabilities to perform the intended task on arbitrary data. Notice that during the entire process, the knowledge remains 'hidden' in a network's adaptable connections, hence the name 'connectionist representation'.

The above training paradigm can be enriched with symbolic knowledge in the following way ('symbolic representation'): prior knowledge about a task (initial domain knowledge) is used to initialize a network before to training. This requires a translation of the information from a symbolic into a connectionist representation. The particular method for converting the symbolic representation of knowledge into its equivalent connectionist representation depends on the kind of symbolic knowledge, the learning task, and the network model used for learning. To date, most efforts are directed towards encoding prior knowledge by programming some network weights to specified values instead of choosing small random values. The programmed weights define a starting point for the search of a solution in weight space. The premise is that a better solution will be found faster compared to starting the search from a random point in weight space. The prior knowledge presumably defines a good starting point in the space of adaptable parameters and leads to faster learning convergence and an explicit inductive bias which focuses a network's attention on relevant input features or favors a desirable connectionist knowledge representation. Examples of this approach include prestructuring of feedforward networks with Boolean concepts (see e.g. [15,46]) and imposing rotation invariance in neural networks for image recognition [3]. We should point out that other types of prior knowledge encoding are possible. For instance, rotation-invariance can also be achieved through training, i.e. presenting examples of rotated objects as inputs to a network. The choice of a network architecture itself represents an implicit use of prior knowledge about an application.

Fidelity of the mapping of the prior knowledge into a network is very important since a network may not be able to take full advantage of poorly encoded prior knowledge or, if the encoding alters the essence of the prior knowledge, the prior knowledge may actually hinder the learning process.

Once a network has succeeded in learning a task as measured by its performance on the training data, it may be useful to extract the learned knowledge. The question arises whether it is possible to extract an adequate symbolic representation of the knowledge learned by a network, i.e. a representation that captures the essence of the learned knowledge. In many cases, the extracted knowledge may only approximate a network's true knowledge; however, it is also possible for the extracted symbolic representation to exceed the accuracy of the knowledge stored in a trained network [30].

For feedforward neural networks, it has been shown that knowledge extracted *during* training can be useful for dynamically adapting a network's topology, i.e. the extracted knowledge can be used to guide the search for a solution.

There are advantages to making effective use of prior knowledge that is common to all learning tasks: (1) the learning performance may lead to faster convergence to a solution, (2) networks trained with hints may generalize better to future examples, and (3) explicit

rules may be used to generate additional training data which are not present in the original data set.

The initialization of feedforward networks with Horn clauses has been the predominant paradigm for prior knowledge in the neural networks community. More recent work has shown how recurrent neural networks can be initialized with prior knowledge about a finite-state process [31]. Other examples of using a domain theory for initializing a feedforward neural network have beeen proposed in the literature [16,46]. Prior knowledge can also be used to alter the objective of the hypothesis search space. TangentProp provides explicit knowledge about the derivatives of the function to be learned [35]; it thus overrides backpropagation's bias toward a smooth interpolation between points with explicit training derivatives. Explanation-based neural networks use previously trained neural networks as initial domain theories, and compute training derivatives from each observed training sample that describes the relevance of each input feature [28]. They are then trained using the TangentProp learning algorithm which minimizes the network output error and the error in network derivatives.

2.2. Applications in medicine

For an overview of neural network applications in medicine, see e.g. [4,36]. For a brief summary of neural network methods applied to clinical diagnosis and medical imaging, see [42]. An overview of other data mining techniques with selected medical applications can be found in [23]. Here, we will briefly discuss some representative medical applications of systems that integrate symbolic knowledge with neural networks.

The KBANN approach has been applied to magnetic resonance spectroscopy (MRS) of normal and cancerous breast tissues [42]. The main characteristics of this application are (1) training data is scarce, (2) there is only very limited overlap between the expert knowledge and the available data, i.e. the expert knowledge gives a poor explanation of the training data, and (3) the hybrid approach significantly outperforms traditional neural network learning without prior knowledge. For further details of this application, we refer to Section 5.

A modified version of the KBANN approach which allows for incremental insertion of domain theories, knowledge refinement, rule extraction and validation has been applied to psychotropic induced coma diagnosis [2]. It is part of an experimental system that offers diagnostic tools via the World Wide Web.

Prior knowledge has been successfully integrated with neural networks in the segmentation of medical images [52]. The prior knowledge consisting of anatomical knowledge about the imaged district, the physical principles of image generation, and the regularities of biological structures was integrated into various neural network architectures.

Neural networks have also been proposed as knowledge acquisition engines for the psychological assessment of heart transplant patients prior to surgery [40]. The neural network discovers through training the most relevant features and combination of features for each diagnosis considered. This acquired knowledge is then interpreted and mapped into symbolic diagnosis descriptors; they are used to guide the reasoning process of case-based systems, to retrieve cases from a case library, and to build explanations.

A hybrid symbolic/neural system for assisting the diagnosis of acidosis and anaemia was proposed in [51]. Preprocessed data is mapped onto a two-dimensional Kohonen self-organizing map. A visualization tool constructs a three-dimensional landscape on the map where valleys correspond to groups of data which belong to the same class and mountains represent boundaries between classes. The extracted rules were found to be quite similar to diagnostic rules found in the medical literature. In addition, the system was able to discover new rules which were subsequently verified by medical experts.

A new pruning and rule extraction algorithm for neural networks trained on Wisconsin Breast Cancer Diagnosis data set was discussed in [38]. Unlike other rule extraction techniques that have been proposed in the literature, this technique does not require neurons to operate near their saturation regions; instead, activation values of hidden neurons are clustered. The author report 95% accuracy of the extracted rules on previously unseen data.

Techniques for extracting knowledge in symbolic form have been applied to the diagnosis of hepatobiliary disorders [20]. The data consisted of continuous values of nine measurements collected from patients. A method that generates piece-wise linear discriminant functions for this dataset were found to be more accurate and concise than the rules generated from the discretized dataset.

In related work, the *functional knowledge transfer* between similar learning tasks has been studied in [40]. Instead of explicitly modeling prior knowledge by assigning weight values prior to training as is done in *representational knowledge transfer*, additional training examples from related learning tasks provide an inductive bias that constrains the hypothesis search. A modified version of the multiple task learning method resulted in a superior diagnostic model for coronary artery disease.

3. KBANN

3.1. Introduction

Prior knowledge can be used to derive an initial hypothesis from which to start the search for a solution. In knowledge-based artificial neural networks, an initial domain theory in the form of propositional rules is used to construct a feedforward neural network [46]. The backpropagation learning algorithm is then used to refine that initial domain theory. KBANN provides an inductive bias which is more likely to generalize as predicted by the initial domain theory; backpropagation provides a generalization bias such that networks are more likely to converge toward a solution with small weights.

We use the method proposed in [45] to illustrate how Horn clauses can be encoded into feedforward networks. Other methods only differ in the way neuron inputs are combined (e.g. [22]). The construction of an initial network is based on the correspondence between entities of the knowledge base and neural networks, respectively. Supporting facts translate into input neurons, intermediate conclusions are modeled as hidden neurons, output neurons represent final conclusions; dependencies are expressed as weighted connections between neurons. The neuron outputs are computed by a sigmoidal function which takes as its argument a weighted sum of inputs.

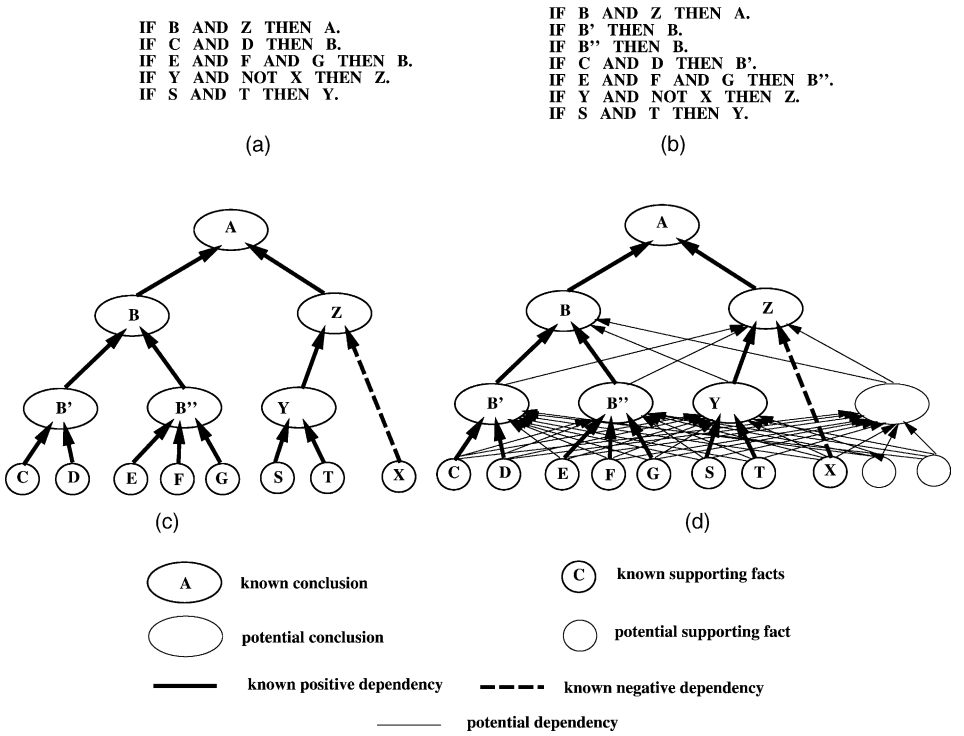


Fig. 2. Construction of KBANNs: (a) original knowledge base, (b) rewritten knowledge base, (c) network constructed from rewritten knowledge base, and (d) network augmented with additional neurons and weights.

Given a set of if–then rules (Fig. 2a), disjunctive rules are rewritten as follows: the consequent of each rule becomes the consequent of a single antecedent; it in turns becomes the consequent of the original rule (Fig. 2b). This rewriting step is necessary in order to prevent combinations of antecedents from activating a neuron when the corresponding conclusion cannot be drawn from such combinations. These rules are then mapped into a network topology as shown in Fig. 2c. A neuron is connected via weight H to a neuron in a higher level if that neuron corresponds to an antecedent of the corresponding conclusion. The weight of that connection is $+H$ if the antecedent is positive; otherwise, the weight is programmed to $-H$. For conjunctive rules, the neuron bias[2] of the corresponding consequent is set to $-(P-(1/2))H$ where P is the number of positive antecedents; for disjunctive rules, the neuron bias is set to $-H/2$. This guarantees that neurons have a high output when all or any one of their antecedents have a high output for conjunctive and disjunctive rules, respectively. If the given initial domain theory is incomplete or incorrect, a network may be supplemented with additional neurons and weights which correspond to rules still to be learned from data (Fig. 2d).

[2] The neuron bias offsets the sigmoidal discriminant function; it is not to be confused with the inductive bias of the learning process.

If an initial domain theory is sparse, the network constructed from the prior knowledge may be too small for a given learning task. In particular, the number of hidden neurons which along with their weights correspond to intermediate conclusions may be insufficient. A heuristic search technique for dynamically creating hidden neurons during the learning process has been proposed [32]. After initial training, a set of tuning examples is used to identify poorly performing hidden units; new hidden units are added as long as a performance improvement can be observed. It has generally been observed that networks initialized with correct prior knowledge train faster and generalize better compared to networks trained without the benefits of an initial domain theory.

3.2. Open question

While good empirical results have been achieved using the framework which combines neural and symbolic learning described above, the merits underlying the symbolic/connectionist approach are not yet well understood. Gaining that insight remains an important open research problem. In this paper, we will address the following open question: how should this explicit inductive bias H be chosen? If we give too little weight to the inductive bias, then it may not be very helpful in finding a solution. If we assign too much importance to it, then the network might not be able to find a solution, particularly when the prior knowledge and the training data do not represent similar concepts.

It is conceivable that the choice of this inductive bias depends on the application, the training data, and the network architecture. By finding a good heuristic for choosing this inductive bias, we can synergistically combine the representational and procedural biases. Bias interactions have also been studied in [7,8]. We proposed a novel heuristic for determining the strength of the inductive bias for feedforward neural networks encoded with prior information using the KBANN method [41]. In Section 4, we will present the details of this heuristic.

4. Strength of inductive bias for KBANN

4.1. Motivation

Based on empirical investigations, the authors of KBANN suggest that all weights which reflect prior knowledge about a learning task be set to $H = 4$. This indiscriminant choice of the inductive bias has two major drawbacks: (1) it is conceivable that different applications require different choices of the inductive bias H which leads to fast convergence and good generalization performance, and (2) it does not provide a mechanism for dealing with uncertainty about the initial domain theory. This section proposes a method for choosing the strength of the inductive bias which takes these two objections into account: the choice of H depends on the application represented by the initial domain theory, the network architecture, and the training data; it adjusts its confidence into the prior knowledge according to the amount and the quality of the available prior knowledge.

Consider an error function E used to train a network. The idea for determining a good value for the inductive bias H is to start the search for a solution at a point in weight space

where the gradient $\partial E/\partial H$ is maximal, i.e. we choose H such that the search starts at a point where the error function in the "direction" of the inductive bias H—the direction of the prior knowledge—is steepest: $\max(|\partial E/\partial H = 0|)$. This avoids the need for determining H through trial-and-error or traversing flat regions of the weight space during the initial training phase. Furthermore, the value H which gives good training performance depends on the prior knowledge and the training data. The function $\partial E/\partial H$ takes both these dependencies into consideration. The more prior knowledge is available and the more accurate that knowledge is, the more the function $\partial E/\partial H$ influences the gradient-descent search for a solution in weight space. Steep descent makes fast convergence possible; furthermore, it is a reasonable premise that good local minima in weight space are more likely to be found at the bottom of steep ravines than in shallow areas.

4.2. Derivation

We will now derive a recursive procedure for evaluating the gradient $\partial E(H)/\partial H$ prior to training which is similar to the error backpropagation learning algorithm. The value of the error function E depends on the particular choice of H, thus $E(H)$. For simplicity, we omit the argument H in the equations for the computation of $\partial E(H)/\partial H$.

Consider the commonly used quadratic error function[3]

$$E(H) = \frac{1}{2}\sum_{p}(d_p - o_p(H))^2 \tag{1}$$

where d_p is the desired network output for pattern p and $o_p(H)$ is the actual network output. Notice that o_p depends on the particular choice of H. For reasons of simplicity, we only consider networks with a single output; the generalization to networks with multiple outputs is straightforward. Then, the derivative $\partial E/\partial H$ is given by

$$\frac{\partial E}{\partial H} = -\sum_{p}(d_p - o_p)\frac{\partial o_p}{\partial H} \tag{2}$$

We can compute $\partial o_p/\partial H$ as follows:

$$\frac{\partial o_p}{\partial H} = o_p(1 - o_p)\sum_{i=1}^{N}\left(\frac{\partial v_i}{\partial H}y_i + v_i\frac{\partial y_i}{\partial H}\right) \tag{3}$$

where $o_p(1 - o_p)$ is the derivative of the sigmoidal discriminant function and v_i is the weight connecting the output of neuron i in the hidden layer immediately preceding the network output layer with the output neuron. The derivative $\partial v_i/\partial H$ can easily be calculated by

$$\frac{\partial v_i}{\partial H} = \begin{cases} +1, & \text{if } v_i = +H \\ -1, & \text{if } v_i = -H \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

[3] Other error functions have also been proposed in the literature. The derivation of the function $\partial E(H)/\partial H$ can be adjusted accordingly.

The derivative $\partial y_i/\partial H$ for neurons in the hidden layers can be computed similarly:

$$\frac{\partial y_i}{\partial H} = y_i(1 - y_i) \sum_{j=1}^{N} \left(\frac{\partial w_{ij}}{\partial H} y_j + w_{ij} \frac{\partial y_j}{\partial H} \right) \tag{5}$$

where $y_i(1 - y_i)$ is the derivative of the sigmoidal discriminant function, w_{ij} connects neuron j with neuron i in the next hidden layer. The derivative $\partial w_{ij}/\partial H$ can easily be calculated by

$$\frac{\partial w_{ij}}{\partial H} = \begin{cases} +1, & \text{if } w_{ij} = +H \\ -1, & \text{if } w_{ij} = -H \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

We need a 'bootstrap' equation in the case where node i is in the first hidden layer, i.e. y_j does not depend on H since it is equal to the value of input neuron j. We then have $\partial y_j/\partial H = 0$ and Eq. (5) simplifies to

$$\frac{\partial y_i}{\partial H} = y_i(1 - y_i) \sum_{j=1}^{N} \frac{\partial w_{ij}}{\partial H} y_j \tag{7}$$

The same equations also apply to the neuron biases.

We have successfully applied this method to a problem in molecular biology [41]. In Section 5, we present a medical application.

5. Application

5.1. ^{31}P MRS of normal breast tissue

Fluctuations in hormone levels during the different phases[4] of the menstrual cycle produce variations in metabolite levels of the breast tissue in women. This well-established observation [42] can be monitored by means of in vivo ^{31}P magnetic resonance spectroscopy. The complexity of the test results requires expert knowledge for their analysis. For a detailed discussion of this complex real-world problem and the knowledge acquisition methods, see [42].

^{31}P MRS is a non-invasive technique for observing phosporus-containing metabolites and intracellular pH. It allows the observation of metabolic activity in cells as it detects the magnetic resonance emitted by cells when exposed to a magnetic field and radio signals. A ^{31}P spectrum of the sampled breast tissue is the result of this method. Peaks in the spectrum correlate to different metabolites (PME, PDE, PCr, Pi, α-ATP, β-ATP and γ-ATP) of the cells. The area under such a peak corresponds to the intensity of the resonance signal for specific nuclei of the cells in the tissue sample. These intensities are used as the data for analyzing the different stages of women's menstrual cycles.

[4] Four menstrual phases: early follicular (ef), late follicular (lf), early luteal (el), late luteal (ll).

Table 1
Training data for magnetic resonance spectroscopy of breast tissue

Volunteer	Phase	Metabolites						
		PDE	PCr	PME	Pi	γ-ATP	α-ATP	β-ATP
2	ef	0.6323	0.1467	0.1588	0.5002	0.1723	0.2115	0.2587
2	lf	0.4324	0.0047	0.2658	0.1763	0.1632	0.2230	0.2074
2	el	0.6133	0.0061	0.1229	0.2855	0.2010	0.2676	0.1649
2	ll	0.6300	0.0799	0.1226	0.3451	0.2200	0.3750	0.2025
3	ef	0.9466	0.0849	0.3191	0.3622	0.3447	0.5459	0.2817
3	lf	0.6604	0.0060	0.2177	0.1159	0.3330	0.2998	0.1959
3	el	0.6429	0.0704	0.0234	0.2234	0.1150	0.2834	0.3028
3	ll	0.9270	0.0077	0.0664	0.3381	0.3571	0.4899	0.2868
4	ef	0.6100	0.0827	0.3381	0.1255	0.2466	0.2817	0.1207
4	lf	0.5504	0.1298	0.1907	0.1723	0.2388	0.3781	0.3413
4	el	0.5660	0.0046	0.0984	0.0972	0.3028	0.3099	0.2811
4	ll	0.3833	0.0941	0.2126	0.0694	0.2453	0.3498	0.1958
5	ef	1.0000	0.0099	0.2012	0.4226	0.2298	0.4669	0.2817
5	lf	0.5651	0.0690	0.2543	0.4362	0.2562	0.2832	0.2167
5	el	0.7325	0.0512	0.1199	0.1711	0.2467	0.1846	0.2740
5	ll	0.6381	0.0060	0.1993	0.2011	0.1993	0.1365	0.1827

Metabolic changes during the four phases of the menstrual cycle. Values correspond to the normalized peak area of seven metabolites extracted from each spectrum.

5.2. Data and initial domain theory

The data[5] contains 16 in vivo ^{31}P MR spectra obtained from four female pre-menopausal volunteers ranging in age from 21 to 45 (see Table 1).

They all had regular menstrual cycles and none were using the contraceptive pill. Four spectra from each volunteer were taken, one at each of the different stages of the menstrual cycle. Seven values were extracted from each spectrum. Each specific normalized value corresponds to a peak area of a specific metabolite present in the spectrum.

The prior knowledge and the resulting KBANN are shown in Fig. 3. We used the real-value encoding of [42] instead of the input encoding method proposed in [46] for the purpose of comparison.

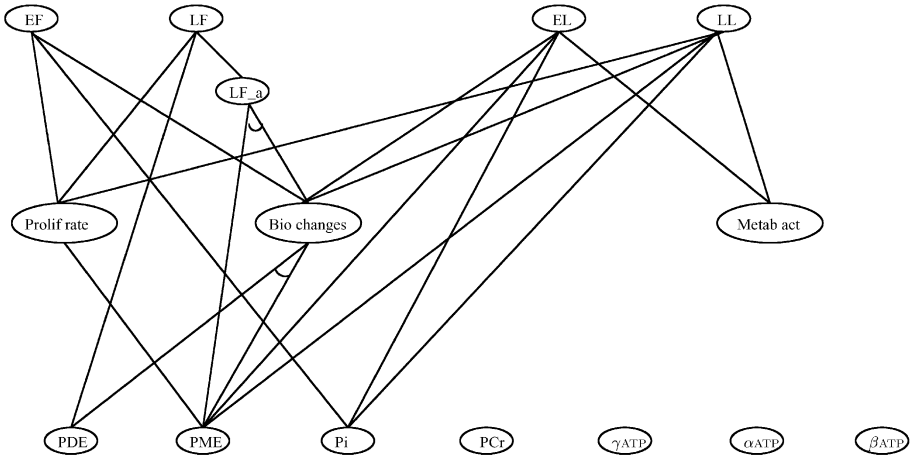
5.3. Training

We performed a four-fold cross-validation on the data. Each fold contained data from three volunteers; the remaining volunteer's data was used for testing. We ran 10 experiments for each fold with different random initialized weights from the interval $[-0.1, 0.1]$. We measured the training and generalization performance for values of H ranging from 0 to

[5] Data provided by the CRC Clinical Magnetic Resonance Research Group, Royal Marsden Hospital, Sutton.

```
ef :- prolif rate                    lf :- prolif rate
ef :- pi level                       el :- metab act
ll :- metab act                      el :- pme level
ll :- pme level                      el :- pi level
ll :- pi level                       ll :- prolif rate
lf :- pde level                      el :- bio changes
ef :- bio changes                    ll :- bio changes
prolif rate :- pme level             lf :- bio changes, pme level
bio changes :- pde level, pme level
```

(a)



(b)

Fig. 3. Prior knowledge: (a) this contains the rules, in PROLOG form, of the knowledge extracted from experts for women's menstrual cycle using ^{31}P MRS, and (b) this is the network structure after the rules have been encoded into the feedforward network according to the KBANN method.

7 in increments of 0.1. All KBANN networks are trained until one of the three following stopping criteria is satisfied:

(1) on 99% of the training examples, the activation of every output unit was within 0.25 of the desired output, or
(2) a network had been trained for 15,000 epochs, or
(3) a network classified at least 90% of the training examples correctly, but had not improved it's ability to classify the training examples for five consecutive epochs.

Table 2
Results of cross-validation

Inductive bias H	Training epochs		Generalisation error	
	μ	σ	μ (%)	σ (%)
$H = 4$	11955	3804	66.3	24.1
Average	8075	1842	72.8	9.6
Optimal training	1198	263	60.6	15.7
$H_{\text{heuristic}}$	1396	235	60.6	12.3

Average and standard deviation are shown for the training time and generalisation performance, respectively, as a function of the inductive bias H for the standard choice $H = 4$, the average over values of H ranging from 0 to 7 in increments of 0.1, the optimal training performance choice, and our heuristic $H_{\text{heuristic}}$.

Neurons had sigmoidal discriminant functions and all networks were trained using the standard quadratic error function (Eq. (1)). A network correctly classified an example if its output was within 0.25 of the desired output. We chose the learning rate $\alpha = 0.5$ and momentum $\beta = 0.7$.[6]

5.4. Results

Fig. 4a–h represent typical simulation results for each of the different folds, respectively. The scarce data for this complex medical domain poses a big challenge. We observe that our heuristic for choosing an explicit inductive bias yields good generalization and training time performance. Variations from fold to fold in training and generalization performance are due to the limited dataset, as for each fold, 25% of the data is set aside for testing.

From the graph of the function $\partial E/\partial H$, we observe that the function $|\partial E/\partial H = 0|$ has a maximum near the inductive bias $H \approx 0.1$. This confirms that the initial domain theory does not fully explain the given training data. A weak inductive bias seems to indicate the programmed network's low confidence in the prior knowledge. We speculate that it is the small training data set and the small overlap between the initial domain theory and the data that leads our heuristic to choose a weak inductive bias. In applications where the initial domain theory and the training data represent similar concepts, we have observed that they have a synergistic effect on the training and generalization performance of neural networks [41].

Average and standard deviation results of the cross-validation for the training and generalisation performances, respectively, are shown in Table 2 as a function of the inductive bias H for the standard choice $H = 4$, the average over values of H ranging from 0 to 7 in increments of 0.1, the optimal training performance choice, and using our heuristic $H_{\text{heuristic}}$ to determine the strength of the inductive bias H.

The initial domain theory only explains 20% of the data. Thus, an average error close to 60% for the cross-validation experiment using our heuristic to encode the networks can be seen as a very good result for this difficult domain. Our heuristic for determining the

[6] These parameters are not necessarily optimal for the networks.

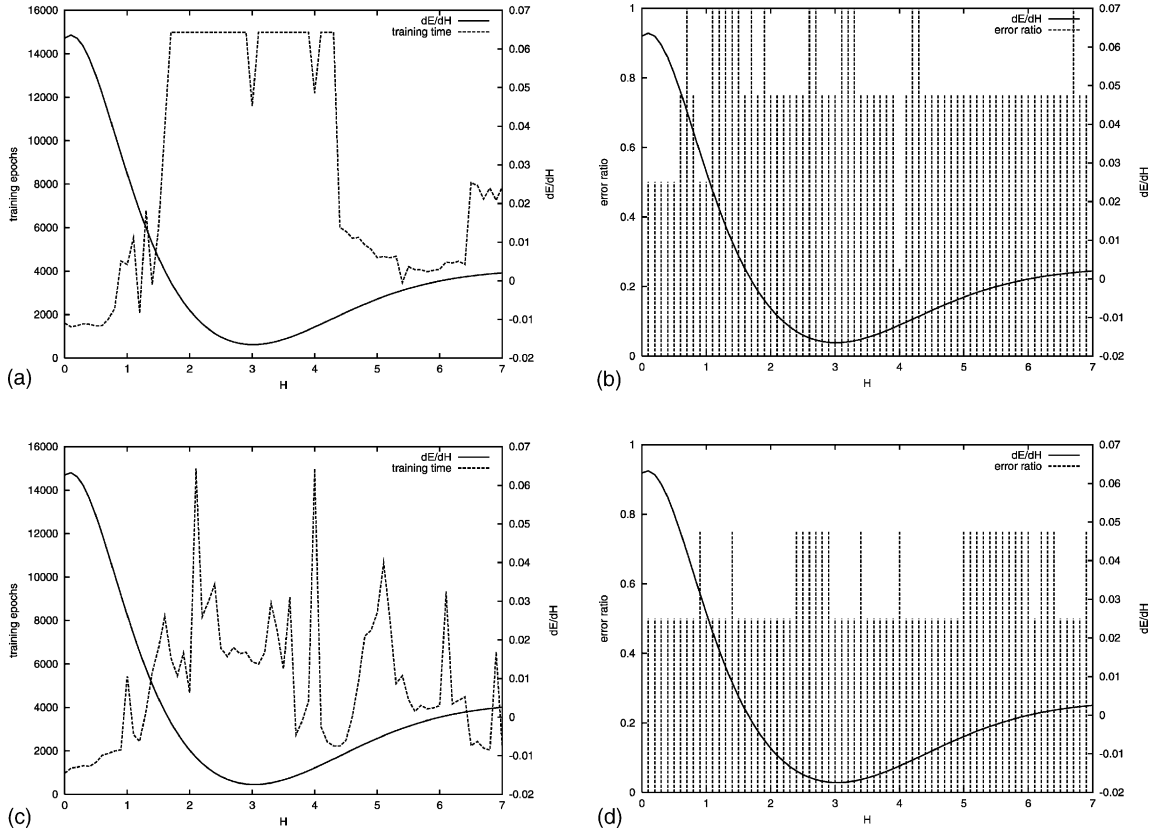


Fig. 4. Cross-validation results: (a), (c), (e), and (g) show typical training times and corresponding generalization performance ((b), (d), (f), (h)) for networks trained with different values of the inductive bias H, for the four different folds, respectively. It plots the function $\partial E/\partial H$ as a function of the inductive bias strength H. Choosing H such that the function $|\partial E/\partial H = 0|$ is maximal results in good performance.

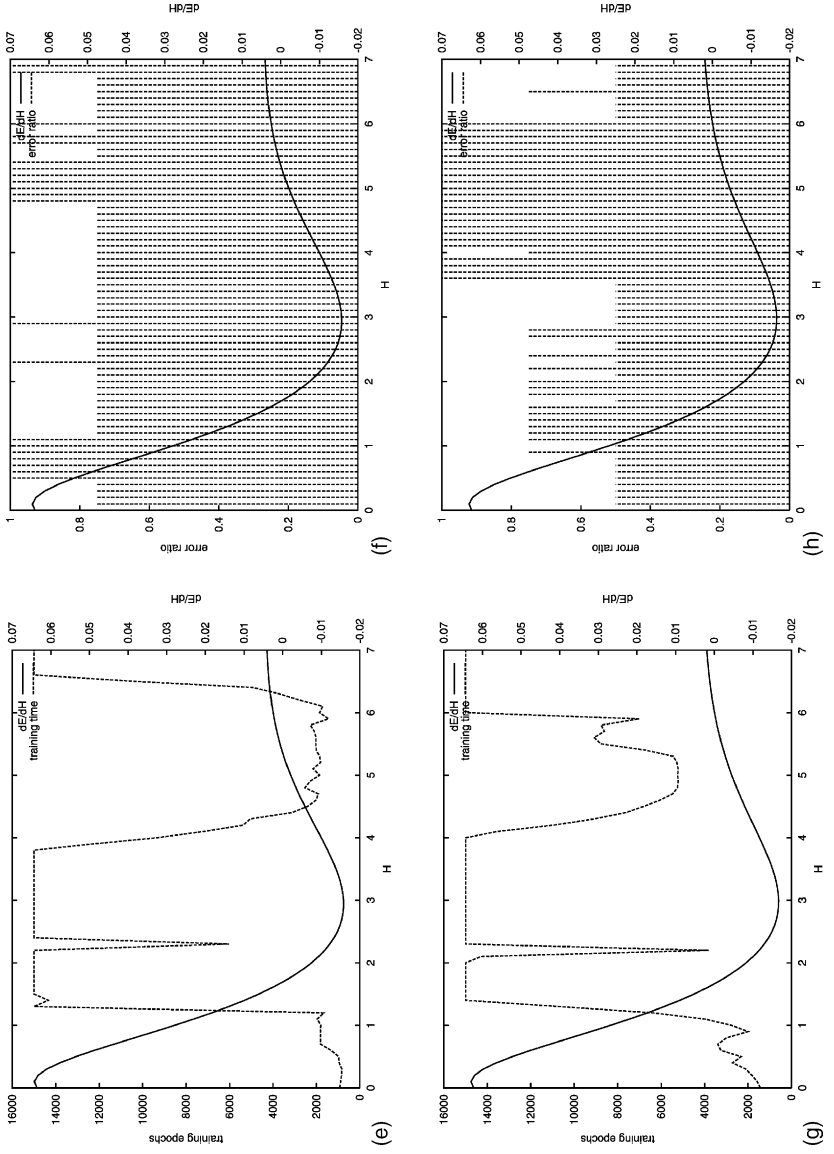


Fig. 4. (*Continued*).

strength of the explicit inductive bias resulted in a 17% improvement of the generalization performance compared to the average choice of the inductive bias. This also exceeds the performance for the standard inductive bias $H = 4$ by 9%.

Our results show a good relative reduction in training time where the inductive bias is chosen according to our heuristic. Training times are reduced by 82% compared to the average choice of the inductive bias H and reduced by 88% compared to the inductive bias $H = 4$; our training times are within 16% of optimal training times. Note that we made no efforts to optimize the training parameters.

6. Conclusions

^{31}P magnetic resonance spectromotry is a typical example of a biomedical application: the obtained results are complex and their interpretation requires expert knowledge. Knowledge-based neural networks have proven useful as they can synergistically combine this expert knowledge with inductive learning from data. The expert knowledge provides an explicit inductive bias for the network training: (1) it determines the network architecture, and (2) instead of initializing all network weights to small random values, it programs weights that correspond to prior knowledge to a value H. This process results in superior training and generalization performance.

In this paper, we have addressed the open question: what strength of the inductive bias H yields good training and generalization performance? We have shown that, for this complex real-world domain, our heuristic for choosing the inductive bias H outperforms the suggested standard inductive bias $H = 4$. Our heuristic chooses the inductive bias H such that the derivative $|\partial E/\partial H = 0|$ of the error function E is maximal. The premises of this heuristic are (1) to start the search for a local minimum in weight space where gradient-descent can rapidly converge to a local minimum, and (2) that good local minima are more likely to be found at the base of deep ravines rather than in shallow areas. Our experiments have shown that our heuristic is not sensitive to the values of initial weights that are not programmed; it takes the initial domain theory, the training data, and the network structure into consideration when choosing a value of H. Thus, our heuristic is able to determine its confidence in the prior information and how well it explains the training data. In this application, the data and the initial domain theory did not represent sufficiently similar concepts; thus, the heuristic chose a low value for the inductive bias H.

We observed statistically significant improvements for both training and generalization performance. Our heuristic reduced the training times by over 80% compared to the average choice of the inductive bias H and the standard choice $H = 4$. The generalization performances improved by 17 and 9%, respectively. Thus, we conclude that our heuristic gives a quantitative measure for successfully dealing with uncertainty in the initial domain theory. Although our results are statistically significant, experiments with more volunteers are neccessary to fully utilize this knowledge for clinical diagnosis. Furthermore, it would be interesting to extract a refined domain theory from trained neural networks in symbolic form and to investigate the accuracy of the extracted rules. Merging complex domain theories into hybrid solutions remains a promising approach to difficult real-world applications.

Acknowledgements

We would like to thank M. Sordo Sánchez, Advanced Computation Laboratory, Imperial Cancer Research Fund, London, and J. Shavlik, Department of Computer Science, University of Wisconsin, for useful discussions.

References

[1] Abu-Mostafa Y. Learning from hints in neural networks. J Complexity 1990;6:192.
[2] Amy B, Danel V, Ertel W, Gonzalez J, Hilario M, Malek M, et al. Modular integration of connectionist and symbolic processing in knowledge-based systems. Tech. rep. W4 D16, final report of ESPRIT project 9119 MIX, Nancy (France): CRIN-INRIA Lorraine; May 1997.
[3] Barnard E, Casasent D. Invariance and neural nets. IEEE Trans Neural Networks 1991;2:498–508.
[4] Baxt W. Application of artificial neural networks to clinical medicine. Lancet 1995;346:1135–8.
[5] Berenji H. Refinement of approximate reasoning-based controllers by reinforcement learning. In: Birnbaum L, Collins G, editors. Proceedings of the Eighth International International Workshop on Machine Learning. San Mateo (CA): Morgan Kaufmann; 1991. p. 475–9.
[6] Brodley CE. Recursive automatic bias selection for classifier construction. Machine Learning J 1995;20:63–94.
[7] Cardie C. Using cognitive biases to guide feature set selection. In: Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society. Bloomington (IN): Lawrence Erlbaum Associates; 1993. p. 469–71.
[8] Cobb H. Inductive biases in a reinforcement learner. In: Proceedings of the ML92 Workshop on Biases in Inductive Learning. San Francisco: Morgan Kaufmann; 1992. p. 1–13.
[9] desJardins M, Gordon DF. Evaluation and selection of biases in machine learning. Machine Learning J 1995;20:1–17.
[10] Frasconi P, Gori M, Maggini M, Soda G. Unified integration of explicit rules and learning by example in recurrent networks. IEEE Trans on Knowledge Data Eng 1995;7(2):340–6.
[11] Fu L, editor. Proceedings of the International Symposium on Integrating Knowledge and Neural Heuristics. Pensacola (FL): University of Florida and American Association for Artificial Intelligence; 1994.
[12] Fu L. Learning capacity and sample complexity on expert networks. IEEE Trans Neural Networks 1996;7(6):1517–20.
[13] Fu L. Integration of neural heuristics into knowledge-based inference. Connection Sci 1989;1:325–40.
[14] Fu L. Rule generation from neural networks. IEEE Trans Systems Man Cybernet 1994;24(8):1114–24.
[15] Fu LM, Fu LC. Mapping rule-based systems into neural architecture. Knowledge-Based Syst 1990;3(1):48–56.
[16] Gallant S. Connectionist expert systems. Commun ACM 1988;31(2):152–69.
[17] Geman S, Bienenstock E, Dourstat R. Neural networks and the bias/variance dilemma. Neural Comput 1992;4(1):1–58.
[18] Gordon D, Perlis D. Explicitly biased generalization. Comput Intell 1989;5(2):67–81.
[19] Hayashi Y. A neural expert system with automated extraction of fuzzy if–then rules and its application to medical diagnosis. In: Lippmann R, Moody J, Touretzky D, editors. Advances in neural information processing systems, vol. 3. San Mateo (CA): Morgan Kaufmann; 1991. p. 578–84.
[20] Hayashi Y, Setiono R, Yoshida K. A comparison between two neural network rule extraction techniques for the diagnosis of hepatobiliary disorders. Artif Intell Med 2000;20(3):205–16.
[21] Kedar-Cabelli S, McCarty T. Explanation-based generalization as resolution theorem proving. In: Proceedings of the Fourth International Workshop on Machine Learning. San Francisco: Morgan Kaufmann; 1987. p. 383–9.
[22] Lacher R, Hruska S, Kuncicky D. Backpropagation learning in expert networks. IEEE Trans Neural Networks 1992;3(1):62–72.

[23] Lavrac N. Selected methods for data mining in medicine. Artif Intell Med 1999;16(3):3–23.
[24] Maclin R, Shavlik J. Refining algorithms with knowledge-based neural networks: improving the Chou–Fasman algorithm for protein folding. In: Hanson S, Drastal G, Rivest R, editors. Computational learning theory and natural learning systems, vol. 1. Cambridge (MA): MIT Press; 1994.
[25] Mahoney J, Moore R. Combining neural and symbolic learning to revise probabilistic rules bases. In: Hanson S, Cowans J, Giles C, editors. Advances in neural information processing systems, vol. 5. San Mateo (CA): Morgan Kaufmann; 1993.
[26] McMillan C, Mozer M, Smolensky P. Rule induction through integrated symbolic and subsymbolic processing. In: Moody J, Hanson S, Lippmann R, editors. Advances in neural information processing systems, vol. 4. San Mateo (CA): Morgan Kaufmann; 1992. p. 969–76.
[27] Minsky M, Papert S. Perceptrons. Cambridge (MA): MIT Press; 1969.
[28] Mitchell T, Thrun S. Explanation-based neural network learning for robot control. In: Hanson JCS, Giles C, editors. Advances in neural information processing systems, vol. 5. San Francisco: Morgan Kaufmann; 1993. p. 287–94.
[29] Omlin C, Giles C. Extraction and insertion of symbolic information in recurrent neural networks. In: Honavar V, Uhr L, editors. Artificial intelligence and neural networks: steps toward principled integration. San Diego (CA): Academic Press; 1994. p. 271–99.
[30] Omlin C, Giles C. Extraction of rules from discrete-time recurrent neural networks. Neural Networks 1996;9(1):41–52.
[31] Omlin C, Giles C. Rule revision with recurrent neural networks. IEEE Trans Knowledge Data Eng 1996;8(1):183–8.
[32] Opitz D, Shavlik J. Dynamically adding symbolically meaningful nodes to knowledge-based neural networks. Knowledge-Based Syst 1996;8:301–11.
[33] Pomerleau D, Gowdy J, Thorpe C. Combining artificial neural networks and symbolic processing for autonomous robot guidance. Eng Appl Artif Intell 1991;4(4):279–85.
[34] Provost FJ, Buchanan BG. Inductive policy: the pragmatics of bias selection. Machine Learning J 1995;20:35–61.
[35] Simard P, Victorri B, LeCun Y, Denker J. TangentProp—a formalism for specifying selected invariances in an adaptive network. In: Moody JE, Hanson SJ, Lippmann RP, editors. Advances in neural information processing systems, vol. 4. San Francisco: Morgan Kaufmann; 1992. p. 895–903.
[36] Reggia J. Neural computation in medicine. Artif Intell Med 1993;5(2):143–57.
[37] Scott R. Artificial intelligence: its use in medical diagnosis. J Nucl Med 1993;34(3):510–4.
[38] Setiono R. Generating concise and accurate classification rules for breast cancer diagnosis. Artif Intell Med 2000;20(3):205–16.
[39] Shavlik J. Combining symbolic and neural learning. Machine Learning J 1994;14:321–31.
[40] Silver D, Mercer R, Hurwitz G. The functional transfer of knowledge for coronary artery disease. Tech. rep., Ont.: Department of Computer Science, Unversity of Western Ontario; 1997.
[41] Snyders S, Omlin CW. What inductive bias gives good neural network training performace? In: Proceedings of the IEEE–INNS–ENNS International Joint Conference on Neural Networks, vol. 3. New York: IEEE Computer Society Press; 2000. p. 445–50.
[42] Sordo Sánchez M. A neurosymbolic approach to the classification of scarce and complex data. Ph.D. thesis, Falmer (Brighton): School of Cognitive and Computing Sciences, University of Sussex; March 1999.
[43] Subramanian D. Shifting vocabulary bias in speedup learning. Machine Learning J 1995;20:155–91.
[44] Suddarth S, Holden A. Symbolic neural systems and the use of hints for developing complex systems. Int J Man–Machine Stud 1991;34:291.
[45] Towell G, Craven M, Shavlik J. Constructive induction using knowledge-based neural networks. In: Birnbaum L, Collins G, editors. Proceedings of the Eighth International Machine Learning Workshop. San Mateo (CA): Morgan Kaufmann; 1990. p. 213.
[46] Towell G, Shavlik J. Knowledge-based artificial neural networks. Artif Intell 70, 119–65.
[47] Towell G, Shavlik J, Noordewier M. Refinement of approximately correct domain theories by knowledge-based neural networks. In: Proceedings of the Eighth National Conference on Artificial Intelligence. San Mateo (CA): Morgan Kaufmann; 1990. p. 861.

[48] Tresp V, Hollatz J, Ahmad S. Network structuring and training using rule-based knowledge. In: Giles C, Hanson S, Cowan J, editors. Advances in neural information processing systems, vol. 5. San Mateo (CA): Morgan Kaufmann; 1993. p. 871–8.
[49] Turney P. Bias and the quantification of stability. Tech. rep., Ottawa (Ont.): Institute for Information Technology, National Research Council Canada; 1994.
[50] Turney P. How to shift bias: lessons from the Baldwin effect. Evolut Comput 1997;4(3):271–95.
[51] Ultsch A, Korus D, Kleine TO. Integration of neural networks and knowledge-based systems in medicine. In: Artificial intelligence in medicine, vol. 934 (of lecture notes on artificial intelligence). Heidelberg: Springer-Verlag; 1995. p. 425–6.
[52] Valli G, Poli R, Cagnoni S, Coppini G. Neural networks and prior knowledge help the segmentation of medical images. J Comput Inform Technol 1998;6(2):117–33.