A GENERALIST-SPECIALIST PARADIGM FOR MULTILAYER NEURAL NETWORKS

Laurent Bochereau and Paul Bourgine

CEMAGREF
BP 121, 92164 Antony, France
Tel. (1) 40 96 61 21 / Fax (1) 40 96 60 36

Abstract : This paper presents a Generalist-Specialist paradigm for building neural networks. The underlying principle is derived from an analogy with medicine. In most cases, a general practitioner or generalist is able to make a sure diagnosis ; however, the generalist sometimes hesitates between several diagnoses and chooses to send his patient to see one or several specialists. This neural network architecture offers many advantages : improvement of the network performances, enhancement of the network interpretation and selection of the critical frontiers between attraction valleys. The paper successively presents the definition of ambiguities matrices for generalist networks, methods for constructing specialists and the functioning of the overall network. The first bid choice when playing bridge gives insights to this approach.

1. Introduction

When performing classification or pattern recognition with a neural network, the challenge is to design the simplest network that will completely learn the training base and perform well when generalizing. This is a difficult problem because one does not know the complexity of the training base beforehand. We thus propose a methodology for building neural networks by incrementally adding specialized neurals networks to an initial general-purpose neural network when it appears necessary.

The analogy with medicine provides insights to this approach : whenever the patient has a common disease, the general practitioner or **generalist** is able to make a sure diagnosis ; however, it often happens that the generalist hesitates between several possible diagnoses and prefers to send his patient to see one or several **specialists**.

One of the advantages of this methodology is the allowance of a precise learning of the frontiers between the attraction valleys of each output unit while maintaining a reasonable size of the overall network. After stating the limitations of the generalist networks (§ 2), we will successively discuss the design of specialist networks (§ 3) and the organization of the overall network (§ 4). An example concerning the first bid choice when playing bridge is then presented (§ 5).

2. Generalists' Errors

We are considering here feed-forward type multilayer networks with complete connections between adjoining layers [4,5,9,10]. The number of hidden layers between the input and output layer is arbitrary. The input-output relationship of each unit (except for those belonging to the input layer) is given by :

$$y = f\,[\,net\,] \qquad \text{where} \quad net = \sum_{i=1}^{n} w_i\, x_i \; - \theta$$

where y represents the output of the unit, n the number of input units, w_i the connection weights and θ the threshold. f is chosen to be the sigmoid function for each unit's output function. We are given a set of examples { X^h, Y^h } where X^h takes its values in the hypercube $\{0,1\}^n$ and Y the hypercube $\{0,1\}^m$. The network calculates a mapping F from $\{0,1\}^n$ to $[0,1]^m$ by using the back propagation algorithm [7,10] that consists of a gradient descent method to modify weights and thresholds in order to minimize the error between the desired output and the output signal of the network.

We now assume that, for each example of the learning base, only one output unit is equal to 1, the other units being equal to 0. This is typical of classification problems and can be performed for other problems by recoding the data. The goal of such a general-purpose network is then to predict the right class for each input vector. This prediction problem is a difficult task and often leads to ambiguous situations (several output units being activated at the same level).

There is an ambiguity unless the prediction for the right answer has a large gap with respect to the other predicted answers. A threshold τ ($\tau \geq 0$) can be chosen as a measure of this ambiguity. Statistics can then be applied in order to build an ambiguity matrix A^τ. This matrix is defined as follows :

$$a_{ij}(\tau) = \left[\Sigma_{\text{learning set}}\ Y_i \ \&\ (y_i - y_j \leq \tau) \right] / \left[\Sigma_{\text{learning set}}\ Y_i \right]$$

where Y_i is the only output equal to 1 for the considered example
y_i , y_j are outputs of the neural network

Each $a_{ij}(\tau)$ can be seen as the probability of ambiguity between classes i and j when class i is the right answer. From the definition, $a_{ii}(\tau) = 1$ for all i. For $\tau = 0$, $a_{ij}(\tau)$ becomes the probability of error between classes i and j when class i is the right answer. The $a_{ij}(\tau)$ are increasing functions with τ and there exists τ_0 such that :

$$\forall\ \tau \geq \tau_0,\quad a_{ij}(\tau) = 1 \quad \forall\ (i, j) \in [1, .., n]^2$$

To each τ corresponds a unique ambiguity matrix A^τ that is going to be used in the next section. The value of τ can also be related to the risk taken by the decision maker. Thus, if one is to make a very serious decision, i.e. stop a power plant, τ needs to be large.

3. Specialists' Construction

After defining an ambiguity threshold and building the corresponding ambiguity matrix, one can design specialized neural networks. The options for defining the necessary specialists are the following :

- ◊ number of output classes to be selected to specify a specialist (2, 3, 2 or 3 ...),
- ◊ maintenance of the generalist's hidden layer,
- ◊ addition of hidden units to the generalist's hidden layer,
- ◊ addition of a hidden layer,
- ◊ degree of interaction between specialists,
- ◊ threshold ε for selecting the groups of output classes.

We decided to create a specialist for each ambiguity, i.e. for each couple of classes i and j such that $a_{ij}(\tau) \neq 0$ as shown on figure 1. The worse case will then lead to design m * (m - 1) / 2 specialists ; however, we think that, in practical cases, only a few specialists will need to be created. If this is not the case, it will always be possible to reduce the number of specialists by imposing the condition $a_{ij}(\tau) \geq \varepsilon > 0$. We then reduce the learning set to the examples corresponding to the selected classes. The connections between the input units and the hidden units are frozen and connections between the output units and the input units are added. Then, learning can be performed in order to derive a better prediction between the selected output classes.

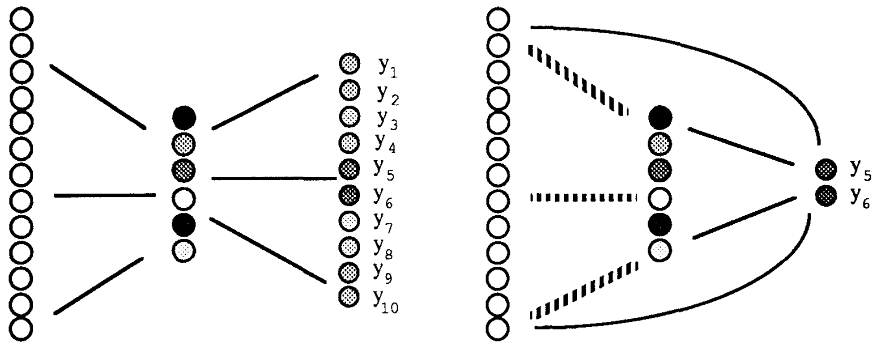

Figure 1 : Construction of a specialist from the generalist

By writing the conditions for which one ouput class is more activated to another one - both also being more activated than a certain level - , one can derive information about the attraction frontiers between both classes. The construction of specialists allows to greatly simplify this task. Furthermore, it is sometimes possible to extract logical and arithmetical rules that would explain the reasoning of the generalist [2,3].

4. Global Architecture

We will now describe a procedure for agregating the individual preferences of the generalist and the specialists. Because of the chosen procedure for constructing the specialists, we have the following rule that governs the relationship between the generalist and the specialists.

Let us call a_0 the preferred decision of the generalist.

Rule : If a preference of the generalist $a_0 > a_i$ is being contradicted by a specialist,

then the specialist's preference is prevalent;

else a_0 is chosen.

This rule implies that the generalist decides unless there exists a specialist that contradicts him. In the other case, the decision is taken among the specialists. The procedure for choosing between the specialists is more explicit when representing the specialists' preferences on a graph G. On the graph G (figure 2), circularities on the preferences may appear. These circularities can be eliminated by taking the reduced graph H which is derived as the quotient of the G graph by the equivalence relation of strong connexity [8]. The corresponding algorithm is linear and the solution(s) are the roots of the reduced graph.

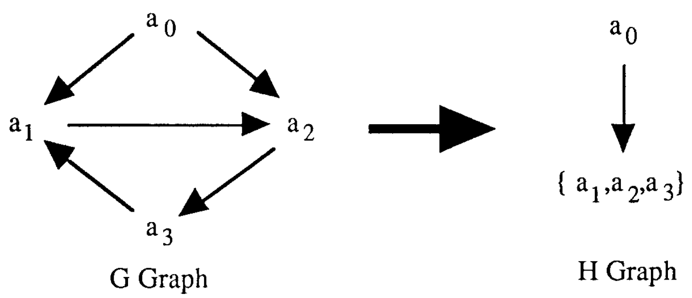

Figure 2 : Reducing the G graph by the equivalence relation of strong connexity

The procedures derived from the preceding rule can then be written as follows :

Let us define :

> Generalist_Preference = a_0
> Specialists_Preferences $= G = \{ a_i, a_j \ / \ a_i > a_j$ and $(i,j) \in \{1,m\}^2 \}$

Then

Procedure *Generalist_vs_Specialists_Choice* (Generalist_Preference,Specialists_Preferences)

> If ($\exists$ i $\in$ {1,m} / ($a_i > a_0$) $\in$ Specialists_Preferences)
> then return (Specialist_Choice (Specialists_Preferences))
> else return Generalist_Preference

Procedure *Specialist_Choice* (Specialists_Preferences)

> H = Strong_Connexity_Reduction (G)
> return (Graph_Roots (H))

5. Application

We considered the problem of the first bid choice when playing bridge [1,6]. A neural network composed of 5 hidden units has been studied in order to produce a generalist. This generalist network gave the right bid for 80 % of the learning base and 66 % of the test base. The ambiguity matrices of the network for $\tau = 0$ and $\tau = 0.05$ are shown below.

$\tau = 0$	P	1 C	1 D	1 H	1 S	1 WT	2 C	2 D	2 H	2 S	2 WT
Pass	1,00	-	-	-	-	-	-	-	-	-	-
1 Club	0,02	1,00	0,10	0,06	0,10	0,10	0,02	0,10	0,06	0,06	0,10
1 Diamond	-	0,02	1,00	0,06	0,02	0,12	0,02	0,30	0,08	0,04	0,14
1 Heart	-	-	-	1,00	-	-	0,02	-	0,22	-	-
1 Spade	0,02	0,02	-	-	1,00	0,02	0,02	0,02	-	0,36	0,06
1 Without Trump	0,02	-	0,02	-	-	1,00	-	0,02	-	-	-
2 Clubs	-	0,08	0,06	0,04	0,04	-	1,00	0,02	0,08	0,08	0,02
2 Diamonds	-	0,06	0,08	0,02	-	0,12	0,02	1,00	0,02	0,02	0,22
2 Hearts	-	-	-	0,40	-	-	0,06	-	1,00	-	-
2 Spades	-	-	-	-	0,08	-	-	-	-	1,00	0,02
2 Without Trump	-	0,02	-	-	-	0,02	-	-	-	-	1,00

$\tau = 0.05$	P	1 C	1 D	1 H	1 S	1 WT	2 C	2 D	2 H	2 S	2 WT
Pass	1,00	0,02	-	-	-	-	-	-	-	-	-
1 Club	0,08	1,00	0,14	0,12	0,12	0,12	0,04	0,12	0,12	0,12	0,12
1 Diamond	-	0,08	1,00	0,12	0,14	0,16	0,02	1,00	0,12	0,10	0,20
1 Heart	-	-	-	1,00	-	-	0,02	-	0,62	-	-
1 Spade	0,02	0,02	-	0,02	1,00	0,02	0,08	0,04	0,02	0,74	0,08
1 Without Trump	0,02	-	0,02	-	-	1,00	-	0,02	-	-	-
2 Clubs	-	0,18	0,06	0,06	0,04	-	1,00	0,02	0,08	0,10	0,10
2 Diamonds	-	0,06	0,88	0,02	0,08	0,18	0,04	1,00	0,04	0,10	0,24
2 Hearts	-	-	-	0,86	-	-	0,12	-	1,00	-	-
2 Spades	-	-	0,02	-	0,50	-	0,02	-	-	1,00	0,02
2 Without Trump	-	0,02	-	-	-	0,06	0,02	-	-	-	1,00

It clearly appears that there is no ambiguity for the "Pass" bid although the "1 Club" and "1 Diamond" bids are difficult to guess correctly. This lead us to develop specialist neural networks in order to validate or modify the predictions of the generalist network. For example, <u>a specialist created to discriminate between 1 Diamond and 2 Diamond was able to learn the reduced training base very quickly and perform with a 100 % success rate</u>. There is a wide range of applications for the Generalist-Specialist paradigm such as the sensory evaluation of food products that is currently under investigation.

6. <u>Conclusion</u>

We have presented a new approach for building neural networks : the Generalist-Specialist paradigm. This approach offers many advantages, such as limiting the network complexity, enhancing the interpretation possibilities, and precisely determining the attraction frontiers between output units. The design of specialists is strongly dependent on the ambiguity threshold. A small threshold will lead to a rather simple network, whereas a larger value ensures a better robustness for the global network. Finally, this design methodology allows one to <u>adjust the architecture's complexity to the overall problem's complexity</u>. .

References

[1] Bochereau L. and Bourgine P., "Implementation and extraction of semantic features from a multilayer neural network : application to the first bid choice when playing bridge", Neuro-Nimes 89, second international worshop, Nimes, 13-16/11/89.

[2] Bochereau L. and Bourgine P., "Extraction of semantic features and logical rules from a multilayer neural network", Proceedings of IJCNN 90, Washington, 15-19/01/90.

[3] Bourgine P., "Propagation et extraction en résolution de problème", Colloque Intelligence Artificielle, Strasbourg, France, 15-19 september, 1986.

[4] Feldman S.E. and Ballard D.H., "Connectionist models and their properties", Cognitive Science, 1982, 6, 205-254.

[5] Fogelman-Soulie F., Gallinari P., Lecun Y., Thiria S., "Automata networks and artificial intelligence" in Automata Networks in Computer Science", F. Fogelman-Soulié, Y. Robert, M. Tchiente eds, Manchester University Press, 1987, 133-186.

[6] Jais P. and Lebel M., La Nouvelle Majeure Cinquième, Editions du rocher, 1982.

[7] Le Cun Y., "A learning procedure for asymmetric networks", Proceedings of Cognitiva 85, Paris, 1985, 599-604.

[8] Minoux M., Gondran M., "Graphes et algorithmes", Eyrolles Eds, 1979.

[9] Personnaz L., Guyon I., Dreyfus G., "Information storage and retrieval in spin-glass like neural networks", J. Physical letters, 1985, 46, L359- L365.

[10] Rumelhart D.E., Hinton G.E., Williams R.J., "Learning internal representations by error propagation" in Parallel Distributed Processing, D.E. Rumelhart & J.L. MacClelland eds, MIT Press, Volume 1: 318-369, 1986.