NEURAL COMPUTING
NCRG
RESEARCH GROUP

Neural Computing Research Group
Dept of Computer Science & Applied Mathematics
Aston University
Birmingham B4 7ET
United Kingdom
Tel: +44 (0)121 333 4631
Fax: +44 (0)121 333 4586
http://www.ncrg.aston.ac.uk/

Combining Spatially Distributed Predictions From Neural Networks

Christopher K. I. Williams
Neural Computing Research Group
Department of Computer Science and Applied Mathematics
Aston University, Birmingham B4 7ET, UK
c.k.i.williams@aston.ac.uk

Technical Report NCRG/97/026 December 1997

Abstract

In this report we discuss the problem of combining spatially-distributed predictions from neural networks. An example of this problem is the prediction of a wind vector-field from remote-sensing data by combining bottom-up predictions (wind vector predictions on a pixel-by-pixel basis) with prior knowledge about wind-field configurations. This task can be achieved using the scaled-likelihood method, which has been used by Morgan and Bourlard (1995) and Smyth (1994), in the context of Hidden Markov modelling.

1 Introduction

Neural networks have been used very successfully in a wide variety of domains for performing classification or regression tasks. A characteristic of most currently successful applications is that the input patterns are either independent (as in static pattern classification) or related over time, rather than being spatially distributed. To extend the use of neural networks to spatially distributed tasks, such as the prediction of a wind vector-field from remote-sensing data, typically it is necessary to combine local bottom-up predictions (wind vector predictions on a pixel-by-pixel basis, based on, e.g. remote sensing observations) with global prior knowledge (typical wind-field configurations, including weather fronts). In this report I show how the prior information and local predictions can be combined using Bayes' theorem to obtain the posterior distribution for the features of interest (the wind-field). This approach is not limited to remote-sensing data, but applies generally to problems where multiple predictions are to be fused with the incorporation of prior knowledge.

2 The generative modelling approach

Suppose we wish to carry out inference on some variables $\boldsymbol{X}$ given some data $\boldsymbol{Y} = \boldsymbol{y}$. In principle the inferential procedure is straightforward; we build a model for $P(\boldsymbol{X}, \boldsymbol{Y})$, and then condition of $\boldsymbol{Y} = \boldsymbol{y}$ to obtain a posterior for $\boldsymbol{X}$. (If the whole posterior is difficult to compute, we may be happy with drawing samples from the posterior, e.g. by Markov Chain Monte Carlo (MCMC) methods.)

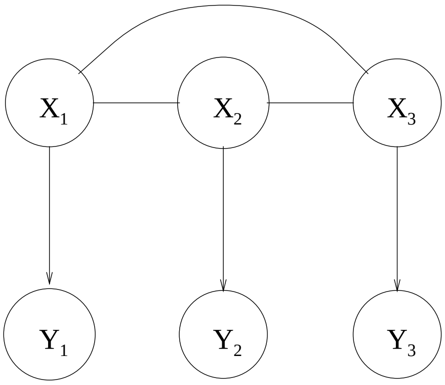

Figure 1: A graphical model illustrating the factorization of $P(\boldsymbol{X}, \boldsymbol{Y}) = P(\boldsymbol{X}) \prod_{i=1}^{k} P(Y_i|X_i)$. Note that the links joining the X nodes are undirected, giving overall a *chain graph* structure.

Of course the joint distribution $P(\boldsymbol{X}, \boldsymbol{Y})$ may be factorized as $P(\boldsymbol{X})P(\boldsymbol{Y}|\boldsymbol{X})$. Under a spatial arrangement of data there may be a component of $\boldsymbol{Y}$ and a component of $\boldsymbol{X}$ at each spatial position. (In the wind field example the Y component is the observations made by a satellite on a region of the sea, and X is the wind vector for that region.) If it is assumed that $P(\boldsymbol{Y}|\boldsymbol{X}) = \prod_{i=1}^{k} P(Y_i|X_i)$, i.e. that the Y observation at each spatial location depends only on the corresponding X, then we obtain

$$P(\boldsymbol{X}|\boldsymbol{Y}) \propto P(\boldsymbol{X}) \prod_{i=1}^{k} P(Y_i|X_i). \tag{1}$$

The factorization of $P(\boldsymbol{Y}|\boldsymbol{X})$ is illustrated as a graphical model in Figure 1.

Note that it is necessary to have a sensor model that gives $P(Y_i|X_i)$ in order to carry out inference. In the next section we present an alternative approach which makes use of predictions of $P(X_i|Y_i)$.

3 The scaled-likelihood method

It may happen that we have a system that predicts $P(X_i|Y_i)$ for each spatial location, and we would like to fuse these predictions with a prior on $\boldsymbol{X}$. The predictions cannot be directly incorporated into equation 1, however, using Bayes' theorem we obtain

$$P(Y_i|X_i) = \frac{P(X_i|Y_i)P(Y_i)}{P(X_i)}. \tag{2}$$

For inference of $\boldsymbol{X}$, the data $\boldsymbol{Y}$ is fixed and so the factor $P(Y_i)$ in equation 2 need not be considered. If we define the *scaled likelihood* (SL) for location i as

$$L_i(Y_i) = \frac{P(X_i|Y_i)}{P(X_i)} \tag{3}$$

then we obtain

$$P(\boldsymbol{X}|\boldsymbol{Y}) \propto P(\boldsymbol{X}) \prod_{i=1}^{k} L_i(Y_i). \tag{4}$$

Thus we have a principled method for the fusion of the predictions $P(X_i|Y_i)$, $i = 1, \ldots k$ with the prior $P(\boldsymbol{X})$. The marginal distribution $P(X_i)$ may be estimated from the data used to train the predictor, perhaps incorporating some other prior knowledge. (Note that

The scaled-likelihood trick has been used by Morgan and Bourlard (1995) and Smyth (1994) in the context of Hidden Markov models, but not, I believe, in the spatial context.

There are a number of issues concerning the merits of the generative and scaled-likelihood approaches:

- The generative approach requires the development a sensor model $P(Y_i|X_i)$. This may need to be quite a complex model, although the predictive distribution $P(X_i|Y_i)$ is actually quite simple, i.e. the generative approach may spend a lot of resources on modelling $P(Y_i|X_i)$ which are not particularly relevant to the task of inferring $\boldsymbol{X}$. On the other hand, having a model for $P(Y_i|X_i)$ is necessary for the detection of outliers.

- Training the models. The scaled-likelihood approach clearly requires the existence of training examples of $\boldsymbol{X}$, decoupling the training of the prior and recognition models. The generative model could be trained using simply $P(\boldsymbol{Y})$, the likelihood of the observations, but if $\boldsymbol{X}$ information is available it would be desirable to use it. For example, the parameters in the model for $P(\boldsymbol{Y}) = \sum_{\boldsymbol{X}} P(\boldsymbol{X})P(\boldsymbol{Y}|\boldsymbol{X})$ may not be identifiable based solely on $\boldsymbol{Y}$ data, but may be if both $\boldsymbol{X}$ and $\boldsymbol{Y}$ data were used.

- Morgan and Bourlard (1995) have used additional contextual information in the prediction of $P(X_i|Y_i)$. In the spatial situation this additional information might be the Y values in spatial locations adjacent to the ith position. Although strictly this would violate the assumptions in equation 1, it may be partially justified using arguments similar to those for the Helmholtz machine (Dayan, Hinton, Neal and Zemel, 1995). In a two-layer Helmholtz machine a feedforward network predicts the posterior distribution for $\boldsymbol{X}$ using the $\boldsymbol{Y}$ values. Our suggestion is to use feedforward networks to approximate the likelihood part of the posterior, and then to combine this term with the prior on $\boldsymbol{X}$ to obtain an approximate posterior.

Finally we note that it is not necessary that the prior distribution of $\boldsymbol{X}$ be represented by a fully connected graph. For example, we are currently experimenting with tree-structured directed acyclic graph models of images (following the work of Bouman and Shapiro (1994)). In this case our prior is over the classifications of pixels (e.g. into road, sky, vegetation etc. classes), and is obtained in an hierarchical manner. Message passing methods (Pearl, 1988) can then be used to find efficiently some properties of the posterior.

Acknowledgements

This work is funded by EPSRC grant GR/L03088, *Combining Spatially Distributed Predictions From Neural Networks.* I thank the Isaac Newton Institute for Mathematical Sciences for its hospitality while this paper was being written, and Chris Bishop, Dan Cornford, Xiaojuan Feng and Ian Nabney for comments on an earlier draft.

References

[1] C. A. Bouman and M. Shapiro. A Multiscale Random Field Model for Bayesian Image Segmentation. *IEEE Transactions on Image Processing*, 3(2):162–177, 1994.

[2] P. Dayan, G. E. Hinton, R. M. Neal, and R. S. Zemel. The Helmholtz Machine. *Neural Computation*, 7(5):889–904, 1995.

[3] N. Morgan and H. A. Bourlard. Neural Networks for Statistical Recognition of Continuous Speech. *Proceedings of the IEEE*, 83(5):742–770, 1995.

[4] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, San Mateo, CA, 1988.

[5] P. Smyth. Hidden Markov models for fault detection in dynamic systems. *Pattern Recogntion*, 27(1):149–164, 1994.