FACTORING NETWORKS BY A STATISTICAL METHOD

Nelson Morgan [†] **and Hervé Bourlard** [‡,†]

[†] International Computer Science Institute, Berkeley, CA 94704, USA

[‡] Lernout & Hauspie Speechproducts, Ieper, B-8900, BELGIUM

We show that it is possible to factor a multi-layered classification network with a large output layer into a number of smaller networks, where the product of the sizes of the output layers equals the size of the original output layer. No assumptions of statistical independence are required.

1 INTRODUCTION

Both on theoretical and practical grounds, it is generally preferable to reduce the number of parameters for a trainable classifier system. In particular, it would be desirable to factor a large Multi-Layer Perceptron (MLP) trained in classification mode (only one output with nonzero output) into two or more smaller ones so that the number of connections can be reduced. In some of these cases, the MLP has an extremely large number of output units, for instance representing a correspondingly large number of pattern types or classes that the net will be trained to recognize. For this case, incorporating a probabilistic interpretation permits a simple factorization of the networks that greatly reduces the number of parameters. No statistical assumptions, such as independence of any of the inputs or outputs, are required. We describe this method, and show an efficient implementation of the resulting networks.

2 MLPs TO ESTIMATE FRAMEWISE PROBABILITIES

Earlier work has shown that the outputs of a Multi-Layer Perceptron (MLP) trained in classification mode with a Least-Mean-Square or relative entropy criterion can be interpreted as posterior probabilities of each class given the input (Bourlard and Wellekens, 1990). In other words, the MLP can estimate the probability $p(q_k|x_n)$ where q_k is a pattern class $\in \mathcal{Q} = \{q_1, \ldots, q_K\}$, the set of all classes for the task (e.g., phonemes for speech applications), and x_n is the input data for pattern n. If there are K classes, then K outputs are required in the MLP. For classification problems with many classes, the corresponding nets would have a large number of connections.

Let coarse categories r_j and s_ℓ be chosen to have a unique correspondence with each class q_k. For instance, r_j and s_ℓ could be the row and column of a matrix of classes q_k. Then,

$$p(q_k|x_n) = p(r_j, s_\ell|x_n) \tag{1}$$

If J is the number of r_j categories and L is the number of s_ℓ categories, then $K = J \times L$. If we use the definition of conditional probability, without any simplifying assumptions the previous expression can be broken down as follows:

$$p(r_j, s_\ell|x_n) \;=\; p(s_\ell|x_n) \times p(r_j|s_\ell, x_n) \tag{2}$$

Thus, the desired probability is the product of one coarse category posterior probability and a second conditional probability. The former can be realized with a standard MLP probability estimator, using the same inputs, and with L output units. Viewing an MLP as an estimator of the left side of a conditional given the right side as input, the second term of (2) can be estimated by an MLP trained to generate the correct class r_j given inputs of class s_ℓ (for instance using one-of-L coding, that is with L inputs only one of which is on) and the input x_n. The latter network has J outputs.

This procedure reduces the training of a single network with $J \times L$ outputs to the training of two smaller networks with J and L outputs respectively, and represents a generic way of splitting large MLPs used in classification mode into two smaller ones. A similar analysis can easily show how the network could be split into three or more smaller ones. If $(J+L) < (J*L)$, the connections to the network outputs are reduced by this procedure. While reducing the number of parameters, this procedure has the potential, however, of requiring much greater computation during the recognition phase. Indeed, if one implements this method naively, the second network must be computed L times for each pattern during recognition, since the output probabilities depend on an assumption of the coarse class s_ℓ (and thus must be evaluated for each such possible hypothesized class). However, this expense can largely be circumvented. Indeed, a simple restriction on the network topology permits the pre-calculation of contributions from hypothesized coarse class s_ℓ to the output; this computation can be done at the end of the training phase, prior to the recognition of any patterns. By simply partitioning the net so that no hidden unit receives input from both s_ℓ input units and data input (x_n) units, we can pre-calculate the contribution to each output unit (prior to any output nonlinearity) for the L possible choices of s_ℓ, and form a $(J \times L)$-table of these contributions. During recognition, the pre-sigmoid output values resulting from data vectors can be computed by a single standard forward pass on the net for each pattern. For each hypothetical pattern category, these contributions from the data inputs can be added to the corresponding context contributions from the table. The major new computation (in comparison with a simple MLP) then is simply the cost of some lookups, particularly for the s_ℓ contributions.

We are currently applying this approach to speaker-independent continuous speech recognition, where it is being used to estimate context-dependent phonetic classes, which could number in the hundreds of thousands (Morgan et al, 1991).

3 DISCUSSION

3.1 The unrestricted split net

When splitting the original big net with $J \times L$ output units into two smaller networks with J and L outputs respectively, the number of parameters is drastically reduced, which could affect the quality of the conditional distributions' estimation. However, parameter reduction is exactly the aim of the proposed approach, both to reduce computation and to improve generalization. As it was done for $p(s_\ell|x_n)$ (Bourlard and Wellekens, 1990; Morgan et al, 1991), it will be necessary to find (e.g. by using cross-validation techniques (Morgan and Bourlard, 1990)) the number of hidden units (and hence the number of parameters) leading to the best estimate of $p(r_j|s_\ell, x_n)$. The desired probabilities can in principle be estimated without any statistical assumptions (e.g., independence). Of course, this is only guaranteed if the training does not get stuck in a local minimum and if there are enough parameters. However, the practical capability of such neural networks to estimate conditional probabilities has already been shown several times (e.g. in the

above references), where x_n in $p(s_\ell|x_n)$ was a vector or a sequence of correlated vectors.

3.2 The topologically restricted net

As shown in the previous section, while reducing the number of parameters, the splitting of the network into two smaller networks results in much greater computation in the contextual network. To avoid this problem it is proposed to restrict the topology of the second network so that no hidden unit shares input from both s_ℓ and x_n . Consequently, the s_ℓ input only changes the output thresholds. However, a recent experiment with frame classification for continuous speech (trained using 160,000 patterns from 500 sentences uttered by a speaker in the Resource Management continuous speech recognition corpus) suggested that this did not affect the correct estimation of $p(r_j|s_\ell, x_n)$. In this example, the network with a split hidden layer predicted (for a test set of 32,000 patterns from 100 sentences) the correct right context 63.6% of the time, while a network with a unified hidden layer predicted the context 63.5% of the time, an equivalent figure.

ACKNOWLEDGEMENTS

Thanks to Chuck Wooters and Steve Renals for running the first tests of this new method, and to our anonymous reviewer for sparking internal discussion about the representation capabilities of our networks.

REFERENCES

Bourlard, H., & Wellekens, C.J., 1990, "Links Between Markov Models and Multilayer Perceptrons", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 12, no. 12, pp. 1167-1178

Morgan, N., Bourlard, H., Wooters, C., Kohn, P., & Cohen, M., 1991, "Phonetic Context in Hybrid HMM/MLP Continuous Speech Recognition", *Proc. of Eurospeech'91*, Genova, pp.109-112

Morgan, N., Bourlard, H., "Generalization and Parameter Estimation in Feedforward Nets: Some Experiments", 1990. *Advances in Neural Information Processing Systems II*, Morgan Kaufmann