ESANN'2000 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2000, D-Facto public., ISBN 2-930307-00-5, pp. 141-146

# Committee formation for reliable and accurate neural prediction in industry

P. J. Edwards and A.F. Murray
Dept. of Electronics and Electrical Eng., Edinburgh University

**Abstract.** This paper describes “cranking”, a new committee formation algorithm. Cranking results in accurate and reliable committee predictions, even when applied to complex industrial tasks. Prediction error estimates are used to rank a pool of models trained on bootstrap data samples. The best are then used to form a committee. This paper presents a comparison of prediction error estimates that may be used for the ranking process. In addition, it shows how the influence of poor models, due to training being unreliable, may be minimised. Experiments are carried out on an artificial task, and a real-world, decision-support task taken from the papermaking industry. In summary, this paper studies committee formation for accurate and reliable neural prediction in industrial tasks.

## 1. Introduction

Model selection techniques may be used for the formation of neural network committees to achieve accuracy and reliability. For multi-layered perceptrons, and other neural network architectures, training is frequently unreliable, finding poor solutions or failing. Prediction error estimation techniques are similarly affected, leading to high predictive loss [2]. These concerns are particularly relevant when neural network algorithms are applied to complex industrial tasks, typically characterised by noisy and sparse data. Such problems may be non-stationary with frequent retraining being necessary. A fully-automated training process is essential for real-time, in-house industrial use of neural technology, as most traditional industries cannot, and will not, employ an engineer with neural expertise. It is therefore vital that the influence of poor quality models be minimised. In this paper we describe *cranking* (committee ranking), a committee formation algorithm that addresses the issue of model reliability for application to complex industrial tasks.

The motivation for this work comes from a specific need for accurate and reliable quality prediction in the papermaking industry. In particular we consider *paper-curl*. Paper-curl is an important quality measure, where bad (high) curl is a major cause of customer dissatisfaction. High curl may result in expensive reprocessing, wasting time, energy and materials. High paper-curl is the major cause of sheet feeding problems in laser printers and photocopiers [11]. In [5, 6] we describe work carried out to develop a neural network predictor of curl. The model may be used to predict curl prior to manufacture, allowing process adjustments to be made. The quantity of paper exhibiting bad curl may thus be reduced, as may wasted plant time, engineering time and energy. This task is non-stationary hence frequent retraining is necessary. In addition, data collection is difficult and expensive leading to data that are sparse and noisy.

The use of a committee of artificial neural network models improves prediction accuracy and reliability [1, 5, 4]. Performance is most improved over that of an individual model if the model outputs are uncorrelated (or "ambiguous") [10]. Output ambiguity, or model diversity can be achieved in a number of ways. The majority of earlier work suggests that varying the constitution of training data sets is most effective [14]. In general, committees provide more accurate and reliable predictive performance than single network predictors. There are intuitive justifications for this. Firstly, neural network training has an inherently stochastic component, as a small change in the data can create a dramatically-different single-network solution. Within a committee the influence of poor models may be minimised. Also, combining networks that exhibit uncorrelated outputs reduces statistical variance in prediction, while leaving bias unchanged [10]. For practical tasks, the use of committees often leads to significant performance improvement. For example, Breiman achieves an average of 20% improvement in performance, applying decision trees to benchmark classification problems [1].

This paper presents a new method for achieving neural network predictions for complex industrial tasks. Predictions are optimised for accuracy and reliability using a method of neural network committee formation called "cranking" (committee ranking). Implicit to this algorithm is the need for model selection and we compare different techniques for prediction error estimation on two tasks, one artificial the other real. Results from experiments using the cranking algorithm are presented and discussed.

## 2. Cranking

In cranking, models are trained using bootstrap subsamples of the training dataset to introduce diversity into an initial population. Each of these models are ranked in terms of their quality based on prediction error estimates. Cranking, or committee ranking, relies upon selecting the best models in terms of estimates of prediction error, from the initial population. By selecting the best models the influence of others, where training has performed poorly or failed, will be minimised. The outputs of the best models are aggregated to form a committee prediction or output.

To rank the models a model selection method is required. In this paper we present a comparison of possible techniques. As each model in the initial population is trained on a bootstrap subsample, out-of-sample patterns are available and these may be used for prediction error estimation, as suggested by Heskes [9]. In comparison, the simplest measure that may be used is the training error, referred to as the Apparent Prediction Error (APE). This generally performs poorly and is included as a control case. We compare this with the Bayesian Information Criterion (BIC) [13], that adjusts the APE to compensate for over-training. These two methods may be used on individual networks. Other more robust techniques that use multiple models are also considered: 10-fold cross-validation (CV(10)) [15]; and the naive and .632 bootstrap [7]. We note here that many other techniques exist. These have been selected as being a representative cross-section of the field, while also being among the more popular techniques. For reviews and more details of these, and other, techniques see the referenced papers and [7, 3].

## 3. Experiments

The first task used in the experiments is an artificial benchmark problem proposed by Friedman [8] and used widely in the machine learning literature. There are five observable variables, $x_1, x_2, \ldots, x_5$ and the task is to model the function :-

$$y = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + \epsilon,$$

where $x_1, x_2, \ldots, x_5$ are generated randomly from uniform distributions [0,1] and $\epsilon$ is distributed normally $N$(0,1). In common with Breiman [1] we use 200 samples for training and use a further 1000 for testing. For this regression task multi-layer perceptrons with four sigmoidal hidden units were used. Training was carried out using approximate Bayesian inference via the evidence framework [12].

The second task is *paper-curl* prediction, as described above. Typical of industrial tasks, this prediction problem is characterised by data that are noisy and sparse. Training is therefore difficult and typically unstable, where a small change in the data leads to a large change in the solution. The use of committees is essential in such applications. In our experiments, eleven parameters related to the manufacture of a roll of paper were used to predict the resulting level of curl, which is measured and predicted on an arbitrary scale of $0<=$ curl $<=90$. 448 data are used for training and 224 as a test set. The regression task was solved via multi-layer perceptron networks with twelve sigmoidal hidden units. Training was again carried out using the evidence framework.

For both tasks training data was sampled via the bootstrap algorithm and the out-of-bag samples were used as a validation set for early stopping. When training via the evidence framework, this is a pragmatic approach as theoretically no early stopping is necessary. However in our experiments, especially for paper-curl prediction, early stopping proved beneficial. We believe this to be attributable to numerical instability in the evidence framework calculations.

## 4. Results

In this section results are presented comparing prediction error estimation techniques for the two tasks, and showing how cranking may be used to reduce the influence of poor models in a committee.

### Comparison of prediction error estimates for committee formation

Experiments were performed to assess the accuracy of prediction error estimates calculated using Heskes' out-of-sample bootstrap technique [9]. Alternative methods were calculated for comparison purposes. They included the APE, the BIC, the CV(10)

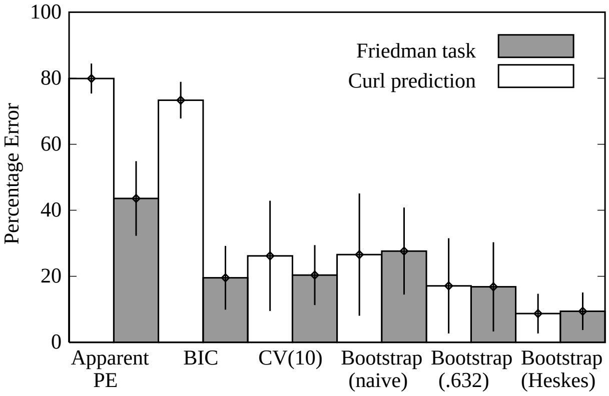

Figure 1: A comparison of different prediction error estimation techniques for use in committee formation. The error bars indicate plus/minus one standard deviation.

measure, the naive bootstrap and the .632 bootstrap. In Heskes' method, a bootstrap data sample is used to train a network, with the out-of-sample data used for early stopping (via a validation set) and also to estimate prediction error. A further 39 networks were then used to adjust this estimate for bias due to the sampling split[1]. Each of the other prediction error estimation methods were applied to the first model (using the same sampled data where necessary). For the naive and .632 bootstrap methods, 40 replicates were used. For the CV(10) measure 4 models were trained on each partition of the data so that a total of 40 networks were used. For the two statistical based methods, the average prediction error estimate over 40 networks was obtained. The results of the experiments are shown in Figure 1, where the percentage error is the difference between the measure in question and the error calculated from a further test dataset.

The results show that while each of the methods perform relatively well for the artificial Friedman task (apart from APE which was included as a worst case control experiment), for real industrial data a more robust approach, such as the CV(10) measure or one of the bootstrap techniques, is necessary. Where out-of-sample data are not available then the .632 bootstrap or 10-fold cross validation may prove effective. For the purpose of committee formation here, however, out-of-sample data are implicitly available and Heskes' bootstrap method makes optimal use of them to provide the most accurate and reliable measures.

### Committee formation using cranking

Experiments were carried out to assess cranking as a method of committee formation. 800 models were trained using bootstrap samples of the dataset. Of these $M$

[1] It should be noted that prediction error estimates were also available for each of the other 39 networks making this method extremely efficient for committee formation.

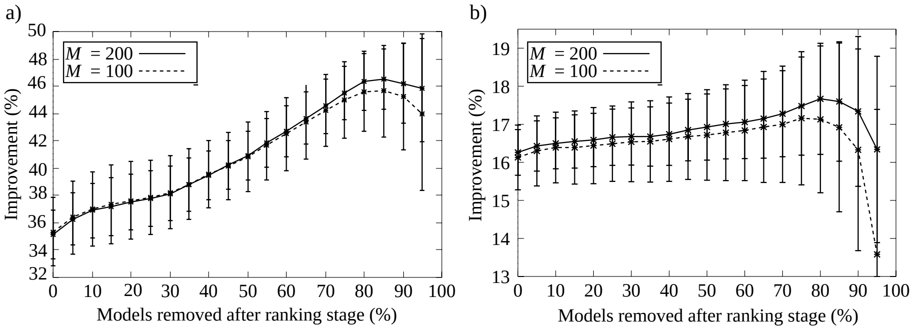

Figure 2: Graphs showing the effect on performance improvement of removing the models with highest prediction error from the committee. The error bars indicate plus/minus one standard deviation. Graph a) shows the results for the Friedman data and graph b) shows the results for paper-curl prediction.

were selected at random and ranked using out-of-sample bootstrap prediction error estimates. Unweighted averaged committees were then formed using the best $C$ of the $M$ networks, where C was increased from 5% to 100% (for $M$=100 and 200). This was repeated 400 times for both the tasks and the results are shown in Figure 2. The results show percentage improvement in performance over that of the average individual model.

The results show that by using only the best 10-20% of the models improved performance may be seen over the case of bagging (where no models are rejected). For the Friedman data the improvement is dramatic (up to 10–11%). For curl prediction the improvement is less marked (up to 1.5–2%), but still significant for this application. For this industrial task training may take place off-line. Therefore with computers becoming increasingly fast, there is little cost in training many (hundreds) of models. The reduction in the number of models used in the final committee is much more important. Delay when waiting for a prediction due to the forward calculation of many networks, and perhaps more significantly the calculation of confidence measures, may be prohibitive. In a longer paper we compare cranking with other committee formation techniques and show that it compares favourably [4]. In summary cranking provides accurate and reliable predictions, results in a small final committee and performs well even when applied to complex industrial tasks.

## 5. Conclusions

In this paper we have presented cranking — a new committee formation method that gives improved predictive performance when training and prediction error estimation are unstable. When multiple models are trained on bootstrap samples of the training data, out-of-sample patterns provide a fast and competitively accurate means of estimating prediction error. These estimates facilitate the ranking of the models in order of their predictive quality. Combining a subset of the better models in an unweighted committee (cranking) reduces the influence of training instability, while also not giving undue significance to the accuracy of prediction error estimates. In summary, cranking is a committee formation algorithm that may be successfully applied to complex industrial tasks, such as paper-curl prediction, where training and prediction error estimation are typically unreliable.

## References

[1] L. Breiman. Bagging predictors. *Machine Learning*, 26(2):123–140, 1996.

[2] L. Breiman. Heuristics of instability and stabilization in model selection. *Annals of Statistics*, 24(6):2350–2383, 1996.

[3] P.J. Edwards and A.F. Murray. A study of early stopping and model selection applied to the papermaking industry. Accepted for publication in the International Journal of Neural Systems, October 1999.

[4] P.J. Edwards, A.F. Murray, and G. Papadopoulos. Cranking : neural network committee formation in the context of high predictive loss. Submitted to the IEEE Transactions on Pattern Analysis and Machine Intelligence, July 1999.

[5] P.J. Edwards, A.F. Murray, G. Papadopoulos, A.R. Wallace, J. Barnard, and G. Smith. The application of neural networks to the papermaking industry. To appear in the IEEE Transactions on Neural Networks, 1999.

[6] P.J. Edwards, A.F. Murray, G. Papadopoulos, A.R. Wallace, J. Barnard, and G. Smith. Paper curl prediction and control using neural networks. *TAPPI Journal*, 82(7):145–151, July 1999.

[7] B. Efron and R.J. Tibshirani. *An introduction to the bootstrap*. Chapman & Hall, New York, 1993.

[8] J. Friedman. Multivariate adaptive regression splines. *Annals of Statistics*, 19:1–141, 1991.

[9] T. Heskes. Balancing between bumping and bagging. In *Proc. Neural Information Processing Systems (NIPS) Conference*, pages 466–472, Cambridge, Massachusetts, 1997. MIT Press.

[10] A. Krogh and J. Vedelsby. Neural network ensembles, cross validation and active learning. In G. Tesauro, D.S. Touretzky, and T.K. Leen, editors, *Proc. Neural Information Processing Systems (NIPS) Conference*, pages 231–238. MIT Press, 1995.

[11] M.B. Lyne. Paper requirements for non-impact. In *International Printing and Graphic Arts Conference Proceedings*, pages 89–97. TAPPI Press, 1988.

[12] D.J.C. MacKay. Bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472, 1992.

[13] G. Schwartz. Estimating the dimension of a model. *Ann. Statist*, 6:461–464, 1978.

[14] A.J.C. Sharkey. On combining artificial neural nets. *Connection Science*, 8(3):299–313, 1996.

[15] M. Stone. Cross-validation : A review. *Math. Operationsforsh. Statist. Ser. Statistics*, 9(1):127–139, 1978.