ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 119-124

Different Criteria for Active Learning in Neural Networks: A Comparative Study

Jan Poland and Andreas Zell

University of Tübingen, WSI - RA
Sand 1, 72076 Tübingen, Germany

Abstract. The field of active learning and optimal query construction in Neural Network training is tightly connected with the design of experiments and its rich theory. Thus there is a large number of active learning strategies and query criteria which have a sound theoretical foundation. This comparative study considers the regression problem of approximating a nonlinear noisy function with relatively few inputs. We evaluate some query criteria, namely space-filling criteria, variance criteria, markov chain monte carlo methods and query by committee.

1 Introduction

An important practical application of Neural Networks is the approximation of an unknown black-box-function with a nonlinear behaviour and not too much inputs. Often the output of the function is corrupted by noise. For example, the function could be given by the behaviour of a highly complex technical system. Then the collection of training data is very expensive with respect to both time and other resources. In this paper, we concentrate on the situation that the bulk of costs is spent for *labelling* the input data, that is *measuring* the output of the black-box-function at a given input $x \in X$, where X denotes the space of valid inputs.

If we assume that it is straightforward to obtain unlabelled input data, for example if the input space is some bounded hypercube in $\mathbb{R}^d$, then we can consider the following problem: Given a training set $(\{x_j\}_{j=1}^n, \{y_j\}_{j=1}^n) \subset X \times \mathbb{R}$ and a trained Neural Model *net*, find a new query $x_{n+1} \in X$ (or a set of new queries) such that the expected information gain is maximal if (x_{n+1}, y_{n+1}) is added to the training set. Here y_{n+1} denotes the result of measuring at x_{n+1}. Then the next query (or queries) can be obtained by optimizing some query criterion on the whole space X. We call this the *optimization method*. In this study, we use a standard nonlinear optimization algorithm, which was started 5 times for each query to prevent it from getting trapped in local optima.

A second scenario arises if there is only a limited set $\Xi = (\{\xi_j\}_{j=1}^k \subset X$ of valid inputs available. This occurs for example when the decision if an input is

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 119-124

is valid or not involves another expensive operation. Alternatively, if the evaluation of the query criterion is very slow but can be performed simultaneously for many points, it might be favourable to generate Ξ at random and use it for constructing the query. Then the next query $x_{n+1} \in \Xi$ is constructed by selecting the point of Ξ which optimizes some query criterion. In this study, this *selection method* is performed with 1000 randomly generated points.

There are very few comparative studies in literature, each of which can cover only a small part of the many different active learning algorithms. This is also true for this study. We concentrate on the situation given above and try to provide some empirical background for the question of which query criteria are good.

2 Different Query Criteria

This section briefly presents and discusses different query criteria. We restrict to the situation that only one point at a time is queried. This has two motivations: First, after adding the query to the training set, the information gained by this step is used in further queries, so this argument suggests that this strategy might carry an improvement. Second, literature observes no measurable drawback of this procedure, see e.g. [2]. A query criterion is then a function $q : (x, \{x_j\}, \{y_j\}, net) \mapsto z \in \mathbb{R}^+$ which takes a new input x, the training set $(\{x_j\}, \{y_j\})$ and the Neural Network *net* and returns some positive measure of the expected information gain if the query is performed at x. Query construction should result in a point x at which q attains its maximum.

a. Random Queries
A simple method (for reference only) is selecting $x \in X$ or $x \in \Xi$ at random.

b. Space-Filling Queries
Another simple and yet powerful strategy is to construct a space-filling design. This can be obtaind by defining $q(x, \{x_j\}) = \min_j \|x - x_j\|$, where $\|\cdot\|$ is a norm on X. This criterion depends neither on $\{y_j\}$ nor on *net*.

c. Network Variance Criteria
Variance criteria emerge from the theory of optimal experimental design. In order to use this theory, several assumptions have to be made:

- the true function is perfectly learnable by the network,
- the weights $w_1 \in \mathbb{R}^W$ of the trained network are close to their optimal (true) value w_0, and this remains valid after adding a new training point,
- the noise in the measurements is independently normally distributed with mean 0 and variance σ^2.

For the weight vector w, the training error function with weight decay parameter α is given by

$$S(w) = \frac{1}{2}\sum_{i=1}^{n}(net(x_i, w) - y_i)^2 + \frac{\alpha}{2}\sum_{i=1}^{W} w_i^2.$$

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 119-124

If we consider the partial derivative of the network output with respect to the weights and the inverse Hessian of the error function,

$$g(x, net) \;=\; \frac{\partial\, net}{\partial w}(x, w_1) \quad \text{and} \quad A(net) \;=\; \Big(\nabla\nabla S(w_1)\Big)^{-1},$$

then the overall predictive variance after a query at x is about minimal if

$$q(x, net) \;=\; g(x, net)^T A(net) g(x, net)$$

is maximal (see [4]). The network weights depend on the training data, therefore this query criterion implicitely depends on the training data, too. Since for linear models the predictive variance at x is given by $g(x)^T A g(x)$ and the second assumption implies that the network is approximately linear in the weights, this criterion maximizes the predictive variance. On the other hand, maximizing $g(x)^T A g(x)$ is equivalent to minimizing $\det(A^*)$ where A^* is the inverse Hessian after adding the new training pattern (x, y). Therefore this criterion is strongly related to the important D-optimality in linear regression.

Theoretical treatments of variance criteria can be found e.g in [4] or [2]. These articles also suggest and derive other and more complex variance criteria, based on thorough probabilistic analysis. But since the dominant term is always $g(x)^T A g(x)$, there is not too much difference in practice.

However, all theoretical analysis does not answer a key question: can we apply these methods if the above assumptions do not hold? Clearly, even for most toy problems not all of them valid. The results of this comparative study suggest that even then the variance criteria are good criteria, at least if the right network architecture is used. For feed-forward networks with sigmoid activation functions of the hidden neurons, and if there are not too few hidden neurons so that the network can capture the uncertainty, the queries obtained by the variance criterion tend to produce a space-filling design as long as the network "is not too sure". In this study, the variance criterion has been used with a single network (denoted by `Var`) and a committee of 5 networks (`Var(C)`).

d. Markov Chain Monte Carlo Methods

An alternative access to the $g(x)^T A g(x)$ criterion considers the distribution of the network outputs. If they are assumed to be normally distributed, a good estimator for their variance (derived in a Bayesian framework) is $\hat{\sigma}^2 + g(x)^T A g(x)$, where $\hat{\sigma}^2$ is the estimated measurement variance (see [1] for a very good overview). So maximizing the query criterion again means finding the point with the maximal predictive variance. Of course the assumption that the distribution of the network outputs is Gaussian holds almost never. Markov Chain Monte Carlo (MCMC) methods (see [3]) provide an alternative technique for estimating the distribution of the network output and thus the predictive variance: The weights are randomly sampled according to their probability (where the probability is given by the training error function), and for each sampled weight the network outputs for a set $\{\xi_j\} \subset X$ of desired input points are calculated. In order to get a representative sample of the true distribution, quite many weights have to be sampled. Therefore, one run takes some time,

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 119-124

and in particular it is prohibitive to use an optimization algorithm that starts the sampling many times. Instead it is convenient use the selection method: generate a number of points in the input space at random, evaluate the MC once for all points and then choose the point with the largest variance. There are several variants of MCMC algorithms. For this study, we used a simple one sampling 1000 weights (denoted by `MCMC`) and a hybrid one sampling 100 weights (`HMC`). Both algorithms are described in [3].

e. Query by Committee
Another query criterion arises when a committee $\{net_j\}_{j=1}^m$ of networks is used as model instead of a single network. The committee members are trained independently with different random initial weights, and the model output at x is the mean of the network outputs $net_j(x)$. The query by committee (QBC, see [5]) criterion is then the committee disagreement or variance, defined by

$$q(x, \{net_j\}) \;=\; \sum_{j=1}^{m} \Big(net_j(x) - \frac{1}{m} \sum_{i=1}^{m} net_i(x) \Big)^2 .$$

It is interesting to note the relation between QBC and MCMC methods: The m different weight vectors of the network committee can be considered as m samples in the weight space according to a probability that is biased towards the peaks. However, if there are only few samples, this bias seems to affect the estimate of the variance in a rather positive than negative way.

For our experiments, we used a committee size of $m = 5$. Moreover, to get a reference, we tried the variance criterion with the committee as well, using $q(x, \{net_j\}) = \sum_j g(x, net_j)^T A(net_j) g(x, net_j)$.

3 Results and Conclusions

Different benchmark functions have been used for this study: the "sine", "hills", "gaussian" and "waves" functions have smooth and moderately complex nonlinear behaviour, the "edge" function has a non-smooth edge and the last function is almost linear. The functions are defined on $[0,1]^d$, where dimensions d from 2 up to 6 have been tested. Although our main interest is the noisy function, we also performed experiments without noise. In all cases, the training set was initialized with $\{x \in \mathbb{R}^d : x_i \in \{0,1\}\} \cup \{x \in \mathbb{R}^d : \exists j \le d : x_j \in \{0,1\},\ x_i = \frac{1}{2} \forall\, i \neq j\} \cup \{\frac{1}{2}\}$ (CCI design). We used feed-forward networks with one hidden tanh layer, linear output, and shortcut connections. We present the results for $2d + 1$ hidden units, other experiments with different hidden layer size and without shortcut connections confirm the results. The networks were trained with a Levenberg-Marquardt algorithm with Bayesian regularization. The figures display the RMSE with respect to the function without noise. All values represent the mean of 20 independent experiments, for clarity of presentation the floating mean over 10 successive numbers of training patterns was taken.

Fig. 1 compares the query criteria with growing number of training patterns for some benchmark functions. The upper two plots show the RMSE

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 119-124

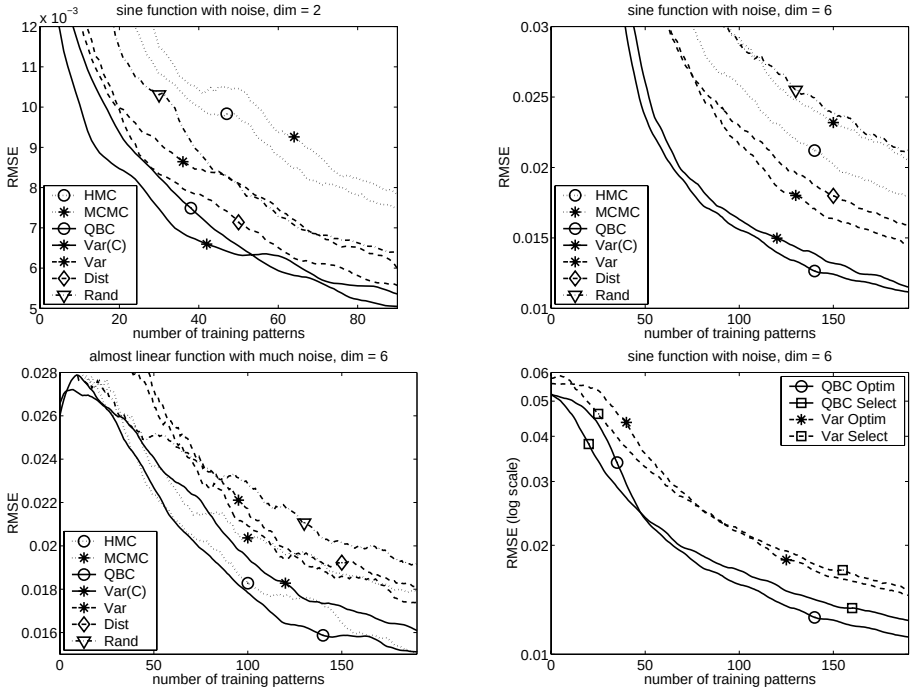

Figure 1: RMSE curves for some of the benchmark functions

development for the sine function in dimension $d = 2$ and $d = 6$. For QBC and the variance criteria the optimization method has been used, for the Markov chain and the distance criteria the selection method. The committee results are always better than the rest. This is not surprising, since network committees are expected to generalize better than a single network. Moreover, the plots indicate that QBC is slightly better than the variance criterion. The Markov Chain methods perform not very well, the random and the space filling criteria deteriorate with increasing dimension.

The curves for the functions without noise are not shown, but they perfectly confirm the results. The lower left plot displaying the curves for the almost linear function in dimension $d = 6$ with a large noise amplitude shows similar results. Here the hybrid Markov chain performs well, but QBC remains the best choice. The lower right plot compares the selection and the optimization methods for the variance criterion (a single network) and QBC. The optimization method results in a lower RMSE finally, while surprisingly the selection method performs better when the number of training patterns is small. Probably the query criteria are highly multimodal in those cases, and the optimization gets stuck in local optima. The other benchmark functions verify the above results, which is shown in fig. 2. Only for the non-smooth

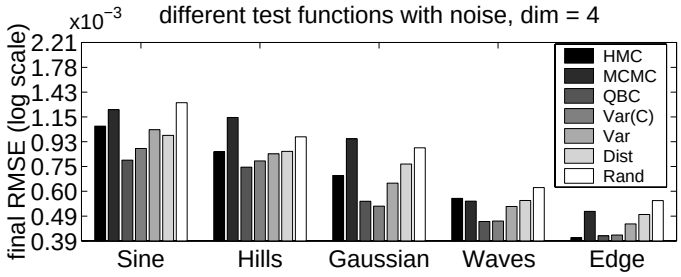

Figure 2: The final RMSE after 150 training patterns

edge function, the HMC criterion performs surprisingly well.

This study was designed and performed in order to increase experience about the practical behaviour of query criteria and thus support the decision for the right active learning strategy. We conclude that if a committee of networks is available, the QBC criterion gives very good results, even with our relative small committee size of 5. Moreover, it is most easy to evaluate, since no Hessian is needed. If only a single network is used, the variance criterion is a good choice. Our space-filling criterion is good in particular in low dimension. The MCMC criteria have a relatively poor performance. Moreover they are computationally very expensive to evaluate, in contrast to the other criteria, the evaluation of which is negligible compared to the network training. We therefore do not recommend MCMC methods for query construction in the setting of this study. However, in high dimensions this may be different.

Acknowledgments. The authors would like to thank Kosmas Knödler, Alexander Mitterer and Thomas Fleischhauer for many helpful discussions. This research has been supported by the BMBF (grant no. 01 IB 805 A/1).

References

[1] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford, 1995.

[2] David A. Cohn. Neural network exploration using optimal experiment design. In J. D. Cowan et al., editor, *Advances in Neural Information Processing Systems*, volume 6, pages 679–686. Morgan Kaufmann, 1994.

[3] R. M. Neal. Probabilistic inference using markov chain monte carlo methods. Technical report, Dept. of Comp. Sc., Univ. of Toronto, Sep 1993.

[4] G. Paass. Query sampling for prediction and model selection (in japanese). *IPSJ Magazine*, 38(7):562–568, 1997.

[5] H. Seung, M. Opper, and H. Sompolinski. Query by committee. In *5th Ann. Workshop on Comp. Learning Theory*, pages 287–294, 1992.