ESANN'2000 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2000, D-Facto public., ISBN 2-930307-00-5, pp. 55-60

# A Statistical Model Selection Strategy Applied to Neural Networks

Joaquín Pizarro Elisa Guerrero Pedro L. Galindo
joaquin.pizarro@uca.es elisa.guerrero@uca.es pedro.galindo@uca.es

Dpto Lenguajes y Sistemas Informáticos e Inteligencia Artificial
Grupo "Sistemas Inteligentes de Computación"
Universidad de Cádiz - SPAIN

## Abstract

In statistical modelling, an investigator must often choose a suitable model among a collection of viable candidates. There is no consensus in the research community on how such a comparative study is performed in a methodologically sound way. The ranking of several methods is usually performed by the use of a selection criterion, which assigns a score to every model based on some underlying statistical principles. The fitted model that is favoured is the one corresponding to the minimum (or the maximum) score. Statistical significance testing can extend this method. However, when enough pairwise tests are performed the multiplicity effect appears which can be taken into account by considering multiple comparison procedures. The existing comparison procedures can roughly be categorized as analytical or resampling based. This paper describes a resampling based multiple comparison technique. This method is illustrated on the estimate of the number of hidden units for feed-forward neural networks.

## 1. Introduction

Many model selection algorithms have been proposed in the literature of various research communities. The existing comparison procedures can roughly be categorized as analytical or resampling based . Analytical approaches require certain assumptions of the underlying statistical model. Resampling based methods involve much more computation, but they remove the risk of making faulty statements due to unsatisfied assumptions [4]. With the computer power currently available, this does not seem to be an obstacle. The standard methods of model selection include classical hypothesis testing, maximum likelihood [2], Bayes method [6], cross-validation [7] and Akaike's information criterion [1].

Although there is active debate within the research community regarding the best method for comparison, statistical model selection is a reasonable approach [5]. We aim at determining which of two models is better on average. A way to define "on average" is to consider the performance of these algorithms averaged over all the training sets that might be drawn from the underlying distribution. Obviously, we have only a limited sample of data, and a direct approach is to divide available data into a training set and a disjoint test set. However, the relative performance can be dependent on the training and test sets.

One way to improve this estimate is to repeatedly partition the data into disjoint training and test sets and to take the mean of the test set errors for these different experiments. The standard t-test for testing the difference between two sample means is not a valid strategy, since the errors are estimated from the same test sample, and are, therefore, highly correlated. A paired sample t-test should be used instead.

However, when more than two models are compared, paired t-tests should be extended to multiple comparison strategies. The first idea that comes to mind is to test each possible difference by a paired t-test. The problem with this approach is that the probability of making at least one Type I error increases with the number of tests made. This phenomenon is called *selection bias*.

A general method to deal with *selection bias* that is useful in most situations is called the Bonferroni multiple comparisons procedure. The Bonferroni approach is a follow-up analysis to the ANOVA method and is based on the following result. If c comparisons are to be made, each with confidence coefficient (1-alpha/c ), then the overall probability of making one or more Type I errors is at most alpha. However, the proper application of the ANOVA procedure requires certain assumptions to be satisfied, i.e., all k populations are approximately normal with equal variances. Residual analysis can be applied to determine whether these assumptions are satisfied to a reasonable degree.

Other procedures, such as Tukey and Tukey-Cramer, may be more powerful in certain sampling situations.

In the following sections, we describe statistical techniques applied to model selection, including significance testing, pairwise comparison and multiple comparison strategies. Then, we justify the use of analysis of variance as a valid strategy to compare different output error means that allows us the estimate of the optimum number of hidden units in feedforward neural networks. Finally, the results of computer simulation for an actual learning task are discussed.

## 2. Strategy description

We will describe our strategy in terms of a classification task by feed-forward neural networks. It is assumed that there exists a set X of possible data points, called the population. There also exists some *target function*, f, that classifies $x \in X$ into one of K classes. Without loss of generality, it is assumed K=2, although none of the results in this paper depend on this assumption, since our only concern will be whether an example is classified correctly or incorrectly. A set of competing models are generated, they differ in the number of hidden units. Misclassifcation errors from the population X is computed for each model and statistical tests are used to decide which of the competing models are better.

Dietterich [3] studied different statistical tests for comparing supervised classification learning algorithms and the sources of variation that a good statistical test should control. In our method, these sources of variation are controlled as follows:

- Selection of the training data and test data. The same training data set and test data set are used to train and test all the competing models. A two-fold cross-validation method is performed since in a k-fold cross-validation method ($k > 2$)

each pair of training sets shares a high ratio of the samples. This overlap may prevent this statistical test from obtaining a good estimate of the amount of variation that would be observed if each training set were completely independent of previous training sets.

- Internal randomness in the learning algorithm. The learning algorithm in each competing model must be executed several times and consequentely several misclassification errors are generated. It is necessary to choose one. If the minimum of these values were taken, this would be the best case and we would think we are near the global minimum of the error function. But this would be a bad selection in a statistical test because an extreme case was chosen. To avoid extreme cases, the maximum and minimum misclassification errors are eliminated and the averaged error is calculated. We are trying to determine *how the model behaves* so we are focusing on the error samples on average better than just considering the minimum error.

Furthermore, we must account the variation from the selection of the test data and from the selection of the training data, so the above process is several times repeated. At the beginning of each iteration, the training and test set are randomly determined. At the end of this process misclassification error mean is calculated. The strategy is summarized as follows:

```
For v:=1 to V  (30 times)
  Random selection of the training and test set, both of them with the same size.
  For h:=model one to model  H
    For  r:=1 to R
          Train model h.
          Error(r) = misclassification error.
    End
    Error_Model(v,h)=Average(Error)
  End;
End;
```

We recommend at least 30 misclassification error samples in order to guarantee the results are distributed according to a normal distribution.

The goal of our strategy is to compare different models and to determine, by analysing the mean and the variance of each one of them, if differences among the models exist. When comparing more than two means, a test of differences is needed . An exploratory/descriptive analysis must be the first step. An univariate analysis of the interval variable by the grouping variable helps to understand the distribution and says whether it is parametric. Both the parametric test for differences (Anova) and the nonparametric test (Kruskal Wallis) for differences are ways to do an analysis of variance. These tests look at how much variation or spread there is in each sub-group. The more within group variation that there is in each sub-group the more difficult it will be to positively say that there is a difference between the group's mean.

There are some questions to be answered:

1- Are the populations distributed according to a Gaussian distribution? While this assumption is not too important with large samples sizes, it is important with small samples sizes (specially with unequal samples sizes). This assumption has

been tested using the method of Kolmogorov and Smirnov and we have always found that the results are according to a Gaussian distribution.

2- Do the populations have the same standard deviations? This assumption is not very important when all the models have the same (or almost the same) number of error subjects, but it is very important when this number differs. In our method the number of error subjects is the same in all the models.

3- Are the data unmatched? We have to compare the differences among group means with the pooled standard deviations of the groups. In our experiment the data are matched.

4- Are the difference between each value and the group mean independent? This assumption is in practice difficult to test. We must think about the experimental design As the sources of variation have been taken into account, we assume this difference is independent.

In our method, the assumptions to use the Anova test have been met. Since a large number of competing models is compared, Bonferroni correction is applied to deal with *selection bias*.

The null hypothesis is usually rejected. In other words, variation among misclassification error means is significantly greater than expected by chance. Thus, groups of models with not significantly different misclassification error means are estimated. To do this, the models are sorted by the misclassification error mean. Two groups are not significantly different if

$$\left|\bar{y}_i - \bar{y}_j\right| \le t_{\alpha/2*c} S_{VNE} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \quad \text{i,j=1,..,M}$$

where M is the max number of models, $n_i$ is the number of data for model i, $\bar{y}_i$ and $\bar{y}_j$ are the means for models $i$ and $j$, $t$ is Student pdf with n-M degree of freedom. c is the Bonferroni correction, $\alpha$ is the statistical significance and

$$S^2_{VNE} = \left(\sum_{i=1}^{M}\sum_{h=1}^{n_i}(y_{ih} - \bar{y}_i)^2\right)/(n_i - M)$$

is the within-sample variation.

In the group with the least misclassification error mean the model with the least hidden units is selected. (Occam's razor criteria).

We have assumed that the goal is to find a network having the best generalization performance. This is usually the most difficult part of any pattern recognition problem, and is the one which typically limits the practical application of neural networks. In some cases, however, other criteria might also be important. For instance, speed of operation on a serial computer will be governed by the size of the network, and we might be prepared to trade some generalization capability in return for a smaller network.

It is desirable to consider a set of several competing models simultaneously, compare them and come to a decision on which to retain. We have therefore been concerned primarily with the choice of a model from a set of competing models rather than with the decision whether or not a new model with more hidden units should be used .

ESANN'2000 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2000, D-Facto public., ISBN 2-930307-00-5, pp. 55-60

## 3. Simulation results

Let us consider the problem of determining the number of hidden units in a feed-forward neural network in a classification task. Let us define a data set where each input vector has been labelled as belonging to one of two classes $C_1$ and $C_2$. Figure 1 shows the input patterns. The sample size is N1=270 data of the class C1 and N2=270 of the class C2.

In the simulation study, we consider multi-layer perceptrons having two layers of weights with full connectivity between adjacent layers. One linear output unit, M sigmoid (logistic, tanh, arctan, etc.) hidden units and no direct input-output connections. The only aspect of the architecture which remains to be specified is the number M of hidden units, and so we train a set of networks (models) having a range of values of M.

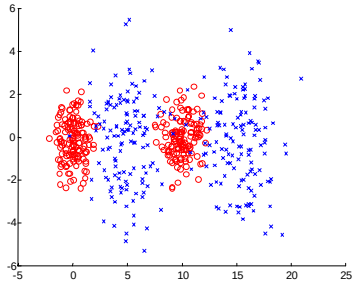


**Figure 1. Sample Data Distribution**

The results of the simulation study are given in Table 2. Two models are in the same group if the difference between its means is less than 0.04973 (statistical significance 0.1). Thus, from the group of models with less error mean (7 hidden units) the model with 4 hidden units could be selected.

**Table 1. Simulation Results**

| Hidden Units | Error Mean | Models not significantly different |
|---|---|---|
| 7 | 0.06139 | 7 6 9 10 8 5 4 |
| 6 | 0.06278 | 7 6 9 10 8 5 4 |
| 9 | 0.06417 | 7 6 9 10 8 5 4 |
| 10 | 0.06546 | 7 6 9 10 8 5 4 |
| 8 | 0.06593 | 7 6 9 10 8 5 4 |
| 5 | 0.07398 | 7 6 9 10 8 5 4 |
| 4 | 0.08630 | 7 6 9 10 8 5 4 |
| 3 | 0.14731 | 3 |
| 1 | 0.27870 | 1 2 |
| 2 | 0.27880 | 1 2 |

If the number of models to compare is increased, results show that four hidden units is a good selection, that is, there is not a statistically significant difference among the error means of neural network architecture with four or more hidden units. The same results are obtained when the number of data is increased.

## 4. Conclusions

An alternative method has been proposed to model selection, where no distribution assumptions about the data are needed. Our goal have been to determine that, in a finite set of models, it is possible to find a subset, whose error mean differences are not significant with respect to the smallest. Our statistical testing procedure has been designed avoiding dependencies and randomness in order to be able to obtain sample data from different models under the same circumstances. After collecting data from a completely randomized design, sample data means are analyzed. The way to determine whether a difference exists between the population means, is to examine the spread (or variation) between the sample means, and to compare it to a measure of variability within the samples. The greater the difference in the variations, the greater will be the evidence to indicate a difference between them. A statistical test procedure has been used to estimate groups of models which differences among the misclassification error means are not significantly greater than expected by chance.

This study shows how statistical methods can be employed for the specification of neural networks architectures. Although the simulation study presented is encouraging, this is only a first step. More experience has to be gained through further simulation with different underlying models, sample sizes and level to noise ratios.

## References

1. H. Akaike, "A New Look at the Statistical Model Identification", *IEEE Transactions on Automatic Contriol*, 1974. AC-19:716-723.

2. C. M. Bishop , *Neural Network for Pattern Recognition*, Clarendon Press- Oxford, 1995.

3. T.G. Dietterich, "Aproximate Statistical Test for Comparing Supervised Classification Learning Algorithms", *Neural Computation*, 1998, Vol. 10, no.7, pp. 1895-1923,.

4. A. Feelders & W. Verkooijen. *"On the statistical Comparison of inductive learning methods"*, Learning from data Artificial Intelligence and Statistics V. Springer-Verlag 1996. pp 271-279.

5. T. Mitchell. *Machine Learning*, WCB/McGraw-Hill, 1997.

6. G. Schwarz, "Estimating the Dimension of a Model", *The Annals of Statistics*, 1978, Vol 6, pp 461-464.

7. M. Stone, "Cross-validatory choice and assesment of statistical prediction (with discussion)". *Journal of the Royal Statistical Society*, 1974, Series B, 36, 111-147.