Statistical Evaluation of Neural Network Experiments: Minimum Requirements and Current Practice*

Arthur Flexer
The Austrian Research Institute for Artificial Intelligence†
Schottengasse 3, A-1010 Vienna, Austria
arthur@ai.univie.ac.at
Technical Report oefai-tr-95-16

Abstract

This work concerns the necessity of statistical evaluation of neural network experiments. This necessity is motivated by applying fundamental notions of statistical hypotheses testing to neural network research. Minimum requirements concerning statistical evaluation are developed and the appropriate statistical techniques are introduced. Articles from two leading neural network journals are examined and critizised for the lack of statistical evaluation they contain.

1 Introduction

There are only few papers that discuss the foundations of the role of experimentation in neural network research, although for the general field of artificial intelligence, recently a whole textbook has been devoted to this problem [Cohen 95]. However, it has already been recognized that the quality of the neural network research practice definitely needs improvement.

[Flexer 95] emphasizes the fact that statistical evaluation is necessary for neural network experiments as for any other empirical science and that problems connected with empirical research and experiment design are wellknown to statisticians, but that there seems to be little awareness of such issues within the neural network community. [Prechelt 96] in his study of 119 articles about neural network learning published in 1993 and 1994 in wellknown journals observes a general lack of comparison with other algorithms and the use of too few and often artificial data sets. 29% of the articles employ not even a single real or at least realistic learning problem. One third of them do not present any quantitative comparison with previously known algortithms at all.

Whereas [Prechelt 96] is concerned with the quantitative amount of evaluation in neural network studies, this paper is concerned with the quality of such evaluations. Minimum requirements for the quality of statistical evaluation will be established that are necessary but maybe sometimes not sufficient, i.e. if an experiment *does not* meet them, its quality will be impaired, if it *does* meet them, there are still other things that can go wrong (e.g. the number of data available for training is too little, the dimensionality of the input vectors is too high, general errors in the design of the experiments, etc.).

2 Why statistical evaluation is a must

We do need experiments in neural network research because the methods we employ and the data we want to analyse are too complex for a complete formal treatment. I.e. for a given data analysis problem we do not have the formal instruments to decide which of the methods is the optimal one. Of course there is a vast literature in statistics, computational learning theory and the like that does help us in such decisions. But the last word in the decision is always spoken by an empirical check, an experiment, as in any other science that needs empirical evaluation of its theories (see [Kibler & Langley 88] for a comparison of physics and machine learning).

The basic structure of neural network experiments is the same as in any other experimental situation: the question is whether there are effects of the variation of the independent variables (mainly type and certain parameter characteristics of the network used and type and characteristics of the data set in question) on the dependent variables (various performance measures like accuracy, root mean squared error or training time).

The observations (in terms of dependent variables) that we make during our experiments are only a portion of the entirety of experiments and observations that are possible in principle. There are at least three

*This work has been published as: Flexer A.: Statistical Evaluation of Neural Network Experiments: Minimum Requirements and Current Practice, in Trappl R., Cybernetics and Systems '96, Proceedings of the 13th European Meeting on Cybernetics and Systems Research. Austrian Society for Cybernetic Studies, Vienna, 2 vols., pp.1005-1008, 1996.

†This work was done in the framework of the BIOMED-1 concerted action ANNDEE ('Enhancement of EEG-Based Diagnosis of Neurological and Psychiatric Disorders by Artificial Neural Networks'), sponsored by the European Commission, DG XII, and the Austrian Federal Ministry of Science, Research, and the Arts, which is also supporting the Austrian Research Institute for Artificial Intelligence.

arguments that motivate this restriction: First, the data available for our experimentation are usually assumed to be just a, hopefully representative, sample of a larger number of data. Second, within such data samples we make divisions into data sets for training and testing, again not all those possible in principle but rather those manageable by our restricted computer sources. Third, there are some random influences (e.g. the random initialization of weights, the sequence in which training data are presented) of which again only a sample can be computed and observed.

But how can we be sure that the portion that we are able to observe is representative of the whole number of events in question? "The procedures of statistical inference" allow us "to draw conclusions from the evidence provided by samples" [Siegel 56]. Only by statistical testing can it be ensured that the observed effects on the dependent variables are caused by the varied independent variables and not by mere chance "that they represent real differences in the larger group from which only a few events were sampled" (ibd.) (i.e. whether the phenomena observed in our sample are significant in a statistical sense or not). Therefore, statistical evaluation of neural network research is in fact a must.

3 Minimum Requirements

Minimum guidelines of proper neural network experimentation can be divided into how to select training and test data and how to statistically evaluate such experiments, whereas the former is a prerequisite for the latter.

3.1 Resampling techniques

Resampling techniques enable it to estimate the performance of a classifier in a fair way, i.e. such that it is guaranteed that approximately the same level of performance will be achieved with a new data set of the same domain.

It is not sufficient to use the so-called *resubstitution* method where the performance of a trained classifier is measured on the data set used for training. It is widely known (even within the neural network or machine learning community, see e.g. [Ripley 92], [Michie et al. 94]) that the performance measure estimated with this resubstitution method is usually over-optimistic, i.e. that the same performance measure computed on new, previously unknown, data is very likely to yield worse results. In statistical terms it is said that such error rates tend to be biased.

Therefore it is at least necessary to *use different sets of data for training and testing*. The simple method of dividing the available data into one training and one test set (2/3 and 1/3 of the data which are mutually exclusive) is called the *holdout* method. Since the neural network cannot use all data for training, performance measures are often pessimistic. Additionally, if such a division into a training and a test set is undertaken, it is necessary to *compute multiple runs* of the experiment in order to avoid random influences (e.g. weight initialization, specific division of the data, sequence of training data). The computation of multiple runs also gives you a better estimate of the true performance.

A simple modification to the holdout method is a *rotation estimator*. The whole data set is divided into K equally sized parts, and each part is used as a test set for a network trained with the remaining data. The observed performance measures for the K different runs are averaged. This procedure is usually known as K-fold *cross-validation*. This technique is still biased in its estimation of performance and there are other techniques like bootstrap [Efron 82] that are able to reduce this bias further at even greater computational cost by using resampling with replacement.

Another important issue, often neglected within neural network research, is the fact that *sometimes another third independent data set is needed* for fair performance estimation. Since it is usually necessary to tune some parameters (e.g. learning rate, number of layers, numbers of units, etc.) to get the best network performance, a division of the available data into three different sets is recommended. [Michie et al. 94] recommend to hold back approximately 20% of the data and divide the remaining data in a set for training and a set for testing, and then tune the parameter using those two sets and an appropriate resampling technique. The final network should use both training and test data for learning with the now optimized parameters and should then be finally tested with the remaining, never before used, 20% of the data. If the use of such a third independent data set is omitted, the obtained error rates will again be biased and over-optimistic because the test set used for repeated tuning in fact becomes a training set. [Mosteller & Tukey 77] (p.37) distinguish between the "form" of a method (i.e architecture of a network including learning parameters) and the "numerical values" of its coefficients (i.e. weight values) and calls the threefold division of data described above "double cross-validation".

3.2 Statistical testing

All statistical tests are only valid under certain conditions and can be divided into parametric and nonparametric methods (see e.g. [Siegel 56]). Parametric methods have a variety of strong assumptions (e.g. that of normal distribution of the data) and are therefore more powerful (i.e. it is easier to come to significant results) than nonparametric methods. From what has been outlined above, it should be clear that multiple runs are necessary for classifier experiments and that usually means over the multiple runs are to be evaluated. The use of parametric methods for the evaluation can be justified with the central-limit theorem which suggests that sample means are normally distributed no matter what distribution the samples themselves form. Therefore, parametric tests can be used for both categorial (e.g. accuracy, sensitivity, specificity) and continuous measures (e.g. root mean squared error, training time) of performance. If parametric tests are being used and the assumption of normality does not hold, it can only happen that instances are being judged as "not significant" that otherwise would have been judged as "significant" but not vice versa (see e.g. [Mosteller & Tukey 77], p.16). If this actually happens, one can still try appropriate nonparametric tests. Therefore, and because of the ease of their computation, we recommend the use of parametric methods like those discussed below as a first approach.

It is by no means justified to report just the best result of the multiple runs of a classifier. Instead, at least *the mean* of the performance measures (e.g. accuracy) over all those runs *and the corresponding variance* σ^2 *should be reported* to give a better estimate of the true performance.

It is even better to *report* the mean over the multiple runs and *the corresponding confidence interval* which can be computed from the standard deviation σ. With a probability of 99% the true value $\bar{X}$ of the observed mean $\bar{x}$ will be within the intervall $\bar{x} \pm 2,58\hat{\sigma}_{\bar{x}}$, with a probability of 95% within $\bar{x} \pm 1,96\hat{\sigma}_{\bar{x}}$, where $\hat{\sigma}_{\bar{x}} = s/\sqrt{N}$ is the standard error estimated from the sample standard deviation s. If the sample is rather small (i.e. number of runs $N << 100$, say less than 30), it is no longer justified to assume normality of the distribution of performance measures. Instead, the distribution forms a Student- or t-distribution and the appropriate t-values have to be used for computation of the confidence intervals which hence become larger.

Confidence intervals allow us to express the amount of uncertainty that comes with every experiment. They also enable use to compare the outcome of experiments under different conditions, e.g. to compare the accuracy means of two different neural networks applied to one data set by computing the confidence intervals for both of them. If the whole intervals do not overlap, there is a statistically significant difference between the two accuracy means. But if only each of the sample means falls outside the confidence intervall around the other mean, a statistically significant difference is not guaranteed.

Therefore it is advised to use a *t-test*, which *should be computed to test the significance of the difference between means* (see [Feelders & Verkooijen 95] or [Egmont-Petersen et al. 94] for a discussion related to neural nets and classifiers in general). The formulas for the computation of the t-test are given in (1), (2) and (3). Assume we have two neural networks A and B and we perform N_{A} runs with A and N_{B} runs with B. $\bar{x}_{\mathrm{A}}$ and $\bar{x}_{\mathrm{B}}$ are the means of the N_{A} and N_{B} runs and s^2_{A} and s^2_{B} are the corresponding variances.

$$t_{\bar{x}_{\mathrm{A}}-\bar{x}_{\mathrm{B}}} = \frac{\bar{x}_{\mathrm{A}} - \bar{x}_{\mathrm{B}}}{\hat{\sigma}_{\bar{x}_{\mathrm{A}}-\bar{x}_{\mathrm{B}}}} \tag{1}$$

$$\hat{\sigma}_{\bar{x}_{\mathrm{A}}-\bar{x}_{\mathrm{B}}} = \sqrt{\hat{\sigma}^2_{pooled}\left(\frac{1}{N_{\mathrm{A}}} + \frac{1}{N_{\mathrm{B}}}\right)} \tag{2}$$

$$\hat{\sigma}^2_{pooled} = \frac{(N_{\mathrm{A}}-1)s^2_{\mathrm{A}} + (N_{\mathrm{B}}-1)s^2_{\mathrm{B}}}{N_{\mathrm{A}} + N_{\mathrm{B}} - 2} \tag{3}$$

We compute the t-value and examine the observed performance difference $\bar{x}_{\mathrm{A}} - \bar{x}_{\mathrm{B}}$ at an appropriate level of significance $\alpha = 0.01$ or 0.05 (i.e. a probabilty of 95 or 99%) and with degrees of freedom $df = N_{\mathrm{A}} + N_{\mathrm{B}} - 2$ for significance with the help of a t-table (for the two-tailed test). Since in the standard comparative experiment the performance measures are all estimated from the same test sample, which makes them highly correlated, a paired sample t-test should be used which gives a more powerful test statistic [Feelders & Verkooijen 95]. This makes it necessary to actually parallelize the samples that are being drawn for neural networks A and B, i.e. to use the same data for training and testing for the networks during the multiple runs. See [Siegel 56] or any standard statistical text book for more details on hypotheses testing and related issues.

It is possible to try to come to significant results by computing more and more runs of an experiment, since higher values for N_{A} and N_{B} imply more degrees of freedom and a decrease of the variance $\hat{\sigma}_{\bar{x}_{\mathrm{A}}-\bar{x}_{\mathrm{B}}}$. But as [Cohen 95] (p.116) points out, this decrease in variance gets rather small when more than 20 runs are being computed.

If more than two means of performances are compared via repeated pairwise t-testing one will end with a high probability to find one or more 'significant' differences when in fact there are none (e.g. for 20 tests with $\alpha = 0.05$, the probability of such an error is 0.64). The simplest approach to deal with this *multiplicity effect* is to divide α through the number of tests that are being performed, which makes it rather hard to come to significant results). Some pointers to more sophisticated solutions can be found in [Feelders & Verkooijen 95] or [Cohen 95] (pp.189).

4 Current Practice

To sum up the previous section, the following can be seen as minimum requirements for proper neural network experimentation:

- the use of different training and test sets
- the computation of multiple runs using an appropriate resampling technique
- the use of a third independent data set in the case of parameter tuning
- to report mean, variance and confidence intervals
- to compute a statistical test (e.g. a t-test) for the comparison of performances

Following the approach in the related study by [Prechelt 96], articles from two leading journals, *Neural Networks* (numbers 1-5 of 1994, Elsevier) and *Neural Computation* (numbers 1-6 of 1994, numbers 1 and 2 of 1995, MIT Press), have been examined as to whether they meet those requirements. Only articles concerned with empirical studies of algorithms applied to practical problems were considered (61 in total).

requirement	yes	no	?
diff. train and test set	72.2%	1.6%	26.2%
multiple runs	57.3%	36.1%	6.6%
$3^r d$ independent data set	4.9%	0.0%	95.1%
mean, var. confidence int.	27.7%	72.3%	0.0%
statistical test	4.9%	93.5%	1.6%

First we want to express our frustration concerning how little care the authors of the examined articles spend on the evaluation of their experiments. Often it is impossible to decide whether the requirements are met by a certain study because the information is simply not in the text. The percentages given in the table above should therefore be seen as very crude but rather over optimistic measurements of the current research practice.

Almost all authors use different data sets for training and testing with the exception of rather questionable theoretical studies, which just want to proove that a certain relation ('exclusive or' is a favourite) is learnable in principle by a certain neural network. The computation of multiple runs seems to be fairly familiar to most authors as well. Because of lack of information, resampling techniques have not been considered explicitly. The use of a third independent data set is only reported in 3 papers (4.9%), for the rest it is often totally unclear if and how parameter tuning was achieved. Only less than one third of the papers do contain computation of mean *and* variance (or confidence intervals) and only 4.9% of them involve a statistical test.

It should be noted that lots of the authors do report means over multiple runs but do not provide standard deviations. Some of the papers contain claims that would need support by empirical simulations but do not include any experimental work at all. Such papers have not been considered in our study and would shift the results even further to the worse.

5 Conclusion

In this work we have motivated the necessity of statistical evaluation of neural network experiments and have given minimum requirements to be met. Our study of articles of two leading journals has shown that concerning our requirements, the quality of the studies in question is rather low. Two possible causes come to mind: Either people working in the field of connectionism are not aware of the necessity of statistical evaluation. Or they are, but are still reluctant to take the consequences since even top journal publications are possible without meeting even the simplest statistical standards of experimentation. In both cases this work should be of help to enhance the quality of connectionist experimentation.

References

[Cohen 95] Cohen P.R.: Empirical Methods for Artificial Intelligence, A Bradford Book, MIT Press, Cambridge, MA, 1995.

[Efron 82] Efron B.: The Jackknife, the Bootstrap, and Other Resampling Plans, Society for Industrial and Applied Mathematics, 1982.

[Egmont-Petersen et al. 94] Egmont-Petersen M., Talmon J.L., Brender J., McNair P.: On the quality of neural net classifiers, Artificial Intelligence in Medicine, 6, 359-381, 1994.

[Feelders & Verkooijen 95] Feelders A., Verkooijen W.: Which method learns most from the data?, Proceedings of the fifth international workshop on AI and Statistics, January 1995, Fort Lauderdale, Florida, pp. 219-225, 1995.

[Flexer 95] Flexer A.: Connectionists and Statisticians, Friends or Foes?, in Mira J. & Sandoval F.(eds.), From Natural to Artificial Neural Computation, Proc.International Workshop on Artificial Neural Networks, Malaga-Torremolinos, Spain, June. Springer, LNCS 930, pp. 454-461, 1995.

[Kibler & Langley 88] Kibler D., Langley P.: Machine Learning as an Experimental Science, Machine Learning, 3(1), 5-8, 1988.

[Michie et al. 94] Michie D., Spiegelhalter D.J., Taylor C.C.(eds.): Machine Learning, Neural and Statistical Classification, Ellis Horwood, England, 1994.

[Mosteller & Tukey 77] Mosteller F., Tukey J.W.: Data Analysis and Regression - a second course in statistics, Addison-Wesley, Reading, MA, 1977.

[Prechelt 96] Prechelt L.: A Quantative Study of Experimental Evaluations of Neural Network Learning Algorithms: Current Research Practice, Neural Networks, Vol. 9, 1996.

[Ripley 92] Ripley B.D.: Statistical Aspects of Neural Networks, Department of Statistics, University of Oxford, 1992.

[Siegel 56] Siegel S.: Nonparametric Statistics for the Behavioral Sciences, McGraw-Hill, Tokyo, 1956.