Some Notes on Neural Learning Algorithm Benchmarking

Lutz Prechelt (prechelt@ira.uka.de)
Fakultät für Informatik
Universität Karlsruhe
D-76128 Karlsruhe, Germany
+49/721/608-4068, Fax: +49/721/694092

Appeared in journal "Neurocomputing", 1995

Abstract

New neural learning algorithms are often benchmarked only poorly. This article gathers some important DOs and DON'Ts for researchers in order to improve on that situation. The essential requirements are (1) Volume: benchmarking has to be broad enough, i.e., must use several problems; (2) Validity: common errors that invalidate the results have to be avoided; (3) Reproducibility: benchmarking has to be documented well enough to be completely reproducible; and (4) Comparability: benchmark results should, if possible, be directly comparable with the results achieved by others using different algorithms.

Keywords: benchmarks, methodology, validity, reproducibility, comparability

1 Introduction

The progress of research in neural network learning algorithms and related issues such as input/output representations etc. is severely slowed down by the bad state of affairs in this area. In most cases benchmarking is not performed with a sufficient number of different problems; rarely can the results presented in articles of different researchers be compared directly; often the benchmark setup is not documented well enough to be reproduced; and in some cases the results are even invalid due to methodological errors during the benchmark process. The purpose of the present article is to make the researchers in the field more aware of these problems and to help avoiding them in the future.

2 Volume

A recent investigation [5] has found benchmarking to be remarkably scarce for neural network learning algorithms, even in journal articles. 34% of all articles presenting a learning algorithm (113 articles were investigated) used zero non-toy learning problems for benchmarking, 41% used but one, and only 6% used more than two!

It is impossible to say how many datasets would be sufficient (in whatever sense) to characterize the behavior of a new algorithm. However, I suggest that at the very least two non-toy learning problems from different domains should be used to benchmark a new algorithm in conference contributions, let alone journal articles. With a smaller number it is impossible to characterize the behavior of a new algorithm in comparison to known ones. The most useful setup is to use both artificial datasets [3], whose characteristics are known exactly, and real datasets, which may have some surprising and very irregular properties. [1] outlines a method for deriving additional artificial datasets from existing real datasets with known characteristics; the method can be used if insufficient amounts of real data are available or if the influence of certain dataset characteristics are to be explored systematically.

3 Validity

Every once in a while articles appear that present benchmarking results that are invalid due to methodological errors. Two of them are most prevailing. First, some researchers use artificial benchmark problems whose structure is known *a priori* to exactly match the structure of the solutions generated by the algorithm, e.g. classify mixtures of Gaussian noise processes with networks of Gaussian basis functions. Such a case hardly ever occurs in a real application and thus the results mean almost nothing.

Second, the results on the test set are often used to adjust parameters of the algorithm such as the network size or a weight decay. Since such a procedure is impossible in a real application, it also invalidates the results. If necessary, a part of the *training* data (called the validation set) has to be set aside for this purpose.

4 Reproducibility

In a majority of cases the information presented in an article about the exact setup of the benchmarking tests is insufficient for other researchers to exactly reproduce it. This violates one

of the most basic requirements for valid experimental science [2]. The most frequent problems are incomplete specification of the values used for the free parameters of the algorithm and the use of training data that nobody else can exactly reproduce.

Parameter vectors should best be specified completely in one place, say, in a figure, and must include all parameters except for initial random weights, for which only the distribution needs to be given. Whenever possible, training data sets should be made available for FTP in the exact form used; the input/output encoding and the partitioning into training, test, and validation data, if any, must be specified painstakingly. FTP availability of data is preferable even for artificial datasets (that could also be represented by their generation rules [3]) in order to avoid human error and stochastic deviations during a reproduction.

5 Comparability

A benchmark is most useful if its results can directly be compared with results obtained by others for other algorithms. However, this is hardly ever the case in neural learning algorithm benchmarking today. Even if two articles use the same data set, the results are most often not directly comparable, because either the input/output encoding or the partitioning of training versus test data is not the same or is even undefined.

We will have to learn to standardize our benchmarking setups so that we can compare benchmarking results just by numbers. Otherwise, the progress in learning algorithms will be slower than necessary. Standardization of course does not mean to fix setups once and forever. Benchmarks *have* to evolve and a voice arguing why some standard setup should be changed will be heard by the community if the argumentation is acceptable. The point is not to change setups unless there is a necessity.

6 The Proben1 collection

In an attempt to help avoid the above-mentioned problems of volume, validity, reproducibility, and comparability, I have prepared a collection of benchmark problem datasets specifically for neural network use, called PROBEN1. This dataset collection[1] is documented in a technical report [4], which also contains advice on how to perform and report benchmarks. This is how PROBEN1 tries to avoid the problems: *Volume* is possible because the collection contains 45 datasets for 15 different learning problems from 12 different domains. All of these problems are what could be called diagnosis problems; 4 are approximation tasks and 11 are classification tasks. The techreport contains some discussion of *validity* to help avoid methodological errors. *Reproducibility* is improved by the standardization of the datasets (including their names, input and output encoding, and the partitioning into training/validation/test data) plus a checklist of items to report about the experiment setup and suggestions on how to report them. *Comparability* is also improved by the standardization of the data plus by suggestions for standard benchmark setups and error normalization terms. A significant body of result data using these setups with several learning algorithms is already available.[2]

It could improve the state of neural algorithm benchmarking if many researchers considered the PROBEN1 techreport while preparing their benchmark setup or a publication reporting the

[1] `ftp://ftp.ira.uka.de/pub/neuron/proben1.tar.gz` or
`ftp://ftp.cs.cmu.edu/afs/cs/project/connect/bench/contrib/prechelt/proben1.tar.gz`

[2] ftp://ftp.ira.uka.de/pub/neuron/nndata.tar.gz

results; even those researchers who do not want to use any of the datasets from the PROBEN1 collection.

7 Science, Publishing, and Progress

A side effect of direct comparability of results will be that researchers as well as reviewers will have to get used to the fact that not every new algorithm can improve on every known one for every learning problem — quite on the contrary! But we should understand ourselves not only as engineers who try to improve, but as scientists who try to understand.

For scientists, however, it should be perfectly acceptable to publish a learning algorithm A based on an idea that was plausible to lead to improvements and was then found not to. If enough results about A are available, such a publication will nevertheless improve our understanding of learning and is thus a perfectly valid scientific contribution.

But as long as we stick to the idea that only improvement counts, researchers will hardly ever be able to make high-volume, reproducible, comparable benchmark results available to the community, because then their articles will usually be rejected due to "lack of progress". But understanding is what counts; and where theory is yet unable to make any exact predictions, benchmarking is a way to gain understanding. This works only, though, when benchmarking is used properly.

Acknowledgements

I thank the editor and the reviewers for their work and comments. Thanks to Scott Fahlman for making PROBEN1 available on his nnbench FTP site.

References

[1] David W. Aha. Generalizing from case studies: A case study. In Derek Sleeman and Peter Edwards, editors, *Machine Learning – Proc. of the 9th Intl. Workshop*, pages 1–10, San Mateo, CA, July 1992. Morgan Kaufman.

[2] Larry B. Christensen. *Experimental Methodology.* Allyn and Bacon, Needham Heights, MA, 6th edition, 1994.

[3] Ray J. Hickey. Artificial universes: Towards a systematic approach to evaluating algorithms which learn from examples. In Derek Sleeman and Peter Edwards, editors, *Machine Learning – Proc. of the 9th Intl. Workshop*, pages 196–205, San Mateo, CA, July 1992. Morgan Kaufman.

[4] Lutz Prechelt. PROBEN1 — A set of benchmarks and benchmarking rules for neural network training algorithms. Technical Report 21/94, Fakultät für Informatik, Universität Karlsruhe, Germany, September 1994. Anonymous FTP: /pub/papers/techreports/1994/1994-21.ps.gz on ftp.ira.uka.de.

[5] Lutz Prechelt. A quantitative study of neural network learning algorithm evaluation practices. In *Proc. 4th Intl. Conf. on Artificial Neural Networks*, Cambridge, UK, June 26-28, 1995. IEE. Anonymous FTP: /pub/papers/techreports/1994/1994-19.ps.gz on ftp.ira.uka.de.