IDEA GROUP PUBLISHING

701 E. Chocolate Avenue, Hershey PA 17033-1117, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

ITB7259

Chapter X

Neural Networks—Their Use and Abuse for Small Data Sets

Denny Meyer
Massey University at Albany, New Zealand

Andrew Balemi and Chris Wearing
Colmar Brunton Research, New Zealand

Neural networks are commonly used for prediction and classification when data sets are large. They have a big advantage over conventional statistical tools in that it is not necessary to assume any mathematical form for the functional relationship between the variables. However, they also have a few associated problems, chief of which are probably the risk of over-parametrization in the absence of P-values, the lack of appropriate diagnostic tools and the difficulties associated with model interpretation. These problems are particularly pertinent in the case of small data sets. This chapter investigates these problems from a statistical perspective in the context of typical market research data.

INTRODUCTION

McCulloch and Pitts (1943) are credited with generating the first interest in neural networks (in the nervous system) but it was not until the 1980s that technology allowed the rapid development of neural networks for solving application problems in numerous fields. The reader is referred to Haykin (1999) for a comprehensive coverage of this evolution.

This chapter appears in the book, Heuristic and Optimization for Knowledge Discovery, edited by Ruhul Sarker, Hussein Abbass and Charles Newton. Copyright © 2002, Idea Group Publishing.

With something akin to horror, statisticians have been watching this recent growth in neural network popularity. Mackinnon and Glick (1999) are particularly concerned by the "black box" or "computational algorithm-oriented" nature of neural network rules. It is difficult to trust a model that is not transparent (i.e., cannot be interpreted). Some of the concern is fuelled by the common (mis)conception that neural networks are about automating data analysis and data modelling (Elder & Pregibon, 1996). Model selection is regarded as a vital part of the statistician's job and to automate this function may seem threatening to statisticians. However, Chatfield (1995) has warned that statisticians have yet to confront the issues surrounding model selection. In particular he points out that the errors caused by model misspecification are likely to be far worse than those arising from other sources. He recommends that statisticians should allow for model uncertainty by averaging over several plausible models or *by choosing a flexible procedure* (such as neural networks) *which does not force a particular form of model on the data.*

Statisticians are being encouraged to test neural networks with typical (small) data sets (Cheng & Titterington, 1994; Warner & Misra, 1996). Although the jury is still out, it seems that, although neural networks need to be applied carefully and creatively (Brierley & Batty, 1999), they have much to offer statisticians as "one of a class of flexible nonlinear regression methods" (Ripley, 1994, p.409). Many studies have focused on a comparison of conventional methods with neural networks (e.g. Cooper, 1999; Haykin, 1999; Ripley, 1996) and several authors have concluded that neural networks should be used in conjunction with conventional statistical tools.

For example, Faraway and Chatfield(1998) suggest that for time series models the Bayesian Information Criterion (Schwarz, 1978) should be used for comparing different models. Borggaard (1995) comments on the advantages of using major principal component scores instead of numerous raw input variables for developing neural network models. Having only a few variables results in smaller networks, which are quicker to train and easier to optimise in a global sense. Markham and Ragsdale (1995) have found that neural networks do not always outperform classical discriminant analysis as a classification tool and advise that a combination of classical and neural network predictions is more accurate. Haykin (1999) recommends the use of cross-validation (Stone, 1974, 1978), to prevent over-parameterisation of neural network models.

From the above it appears that standard statistical tools can be used to improve neural network models. In this chapter we explore this idea further while trying to fit three neural network models using small data sets and typical commercial software. In these models we confront the problems of over-parameterisation, diagnostics and interpretation for small data sets, suggesting how the range and power of commercial neural network software can be enhanced. But this chapter makes no attempt to review the many exciting theoretical developments in this field. Readers are referred to the excellent books of Haykin (1999) and Bishop (1995) for this purpose.

Related Content

Measuring Diversity at a Historically Black College of Dentistry

Garnett Lee Henley, Wanda Lawrence, Candace Mitchell, Donna Henley-Jacksonand Tawana Feimster (2012). *Cases on Institutional Research Systems (pp. 212-227).*

www.irma-international.org/chapter/measuring-diversity-historically-black-college/60849

Incorporating Correlations among Gene Ontology Terms into Predicting Protein Functions

Pingzhao Hu, Hui Jiangand Andrew Emili (2011). *Ontology Learning and Knowledge Discovery Using the Web: Challenges and Recent Advances (pp. 154-173).*

www.irma-international.org/chapter/incorporating-correlations-among-gene-ontology/53885

Deploying Data Warehouses in Grids with Efficiency and Availability

Rogério Luís de Carvalho Costaand Pedro Furtado (2010). *Complex Data Warehousing and Knowledge Discovery for Advanced Retrieval Development: Innovative Methods and Applications (pp. 208-229).*

www.irma-international.org/chapter/deploying-data-warehouses-grids-efficiency/39593

A Case Study: Closing the Assessment Loop with Program and Institutional Data

Robert Elliott (2012). *Cases on Institutional Research Systems (pp. 326-338).*

www.irma-international.org/chapter/case-study-closing-assessment-loop/60858

The Quest for Economic Recovery: Innovative Development and KM Perspectives

Mariza Tsakalerouand Rongbin W. B. Lee (2015). *Knowledge Management for Competitive Advantage During Economic Crisis (pp. 242-249).*

www.irma-international.org/chapter/the-quest-for-economic-recovery/117851