The Use of Boundary Conditions for Inductive Models

Peter Whigham

Spatial Information Research Centre
University of Otago. Dunedin, New Zealand
Phone: +64 3 479-7391 Fax: +64 3 479-8311
Email: pwhigham@infoscience.otago.ac.nz

Presented at SIRC 2004 – The 16th Annual Colloquium of the Spatial Information Research Centre
University of Otago, Dunedin, New Zealand
November 29th-30th 2004

ABSTRACT

There is a large amount of interest in creating models from data using a variety of machine learning methods. Most of these approaches require a good distribution of observed values to produce reliable models. The use of background knowledge to augment the observed values has also been explored as a method to supplement the original feature set of training data. This paper argues that there is an additional set of data that can be created for many types of problems, based on the concept of boundary conditions. This boundary data incorporates an understanding of the modeled system behaviour under certain extreme values and therefore reduces the degrees of freedom within the inferred model. This paper argues that by using this information when training an inductive model a more robust generalization of the data can be achieved under some circumstances.

***Keywords and phrases*:** induction, background knowledge, boundary conditions

1.0 INTRODUCTION

There are a large number of methods designed to construct models based on a set of training data, including regression techniques, neural networks, decision trees, evolutionary computation and inductive logic programming. All of these methods assume that the training data represents a reasonable distribution of the observed data ranges and combinations of possible data that are likely to be observed. Given the observed data the aim is to construct a good generalized model that will perform well on unseen (independent) data that has been drawn from a similar problem and distribution (Langley 1986).

Generally when applying these methods an appropriate set of features that characterize important aspects of the problem are required for the models to perform well. The concept of automatic feature construction has been examined for the past twenty years, with the advent of a number of systems that alter their language to aid the induction process (Lenat 1984; Utgoff 1986; Rendell & Cho 1990; Muggleton & Buntine 1992). Although these approaches have shown some promise, the majority of techniques currently employed rely on the user to construct the features prior to learning (Hirsh & Noordewier 1994). Knowledge of the importance of certain features, usually constructed by combining values from the observed data or as additional measures of certain qualities of the data, has been used successfully to improve model construction (Silvert & Baptist 1998). These approaches can be described as preprocessing of the data to construct relevant features and (in some cases) to reduce the dimensionality of the problem description.

This paper will describe a simple concept to increase the number of training examples used for inductive model construction (supervised learning) that relies on knowledge of the boundary conditions for the system under consideration. It will be argued that for many problem domains there are known values for some boundary conditions, and incorporating these values into the training data will reduce the degrees of freedom of the inferred model, and therefore increase the generalization ability of the final models. This approach is

independent of a particular learning method, and can be used for all forms of data. This paper is structured as follows: §2 will give a motivational example of the use of boundary conditions for a simple model, §3 will demonstrate the use on an ecological time-series model, §4 will discuss the application to a variety of spatial models, and §5 will draw some final comments and conclusions.

2.0 A MOTIVATIONAL EXAMPLE

A simple block data set from the Donoho-Johnston benchmark data (Sarle 1999) has been selected for this example. The inductive model used for all experiments was a simple neural network, Neuroet (Tribou & Nobel 2004). The complete block data set is shown in Figure 1.

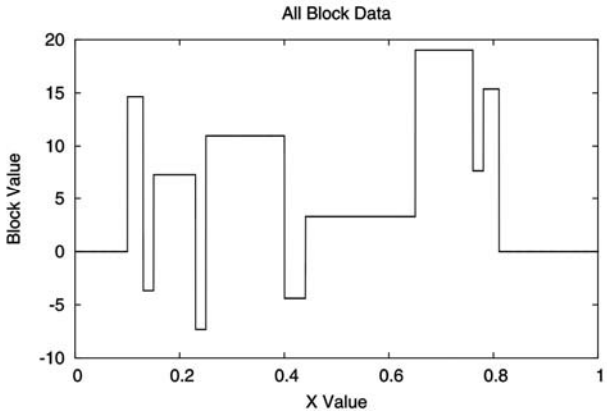

Figure 1. The complete block data set

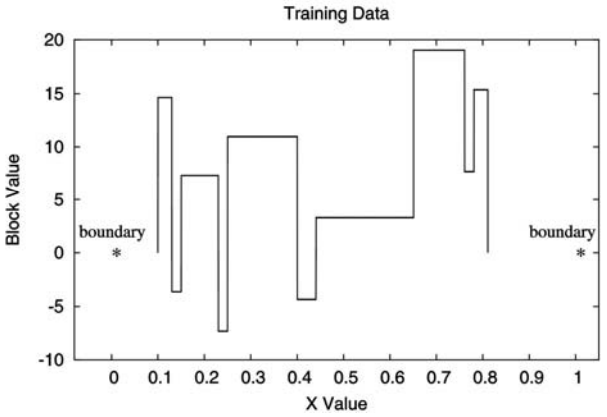

Figure 2. The Blocks training data and boundary conditions

This dataset is non-linear, with boundary conditions of zero at both extremes of the data. The original data was modified so that the start and end portions with value zero were removed. A neural network was then trained on this modified dataset, with the resulting test error given by comparing the model performance applied to the original data. There were 10 models constructed, with a resulting R^2 value of 0.17 ± 0.16. To demonstrate the use of boundary conditions, the modified training dataset was augmented with 100 entries at the X values of 0 and 4096, each with a block value of 0. The training data, plus the location of the two boundary condition points, are shown in Figure 2. The resulting model, applied to the original data set had a significantly improved R^2 value of 0.68 ± 0.25. Although this example is biased, since the values outside of the training data are equal to the boundary value, it shows that holding a function at an endpoint can significantly change the behaviour of the resulting model. Figure 3 shows the best result when trained with and without the boundary condition. In addition, the worst result with the boundary training data is also shown. The main point to note is that due to the boundary condition the resulting neural network has learnt to pull the model back to zero when X is 0 and 1. Since the boundary conditions represent fixed values that the model should correspond to it has changed the

overall shape of the model within the neighbourhood of the boundary, and therefore resulted in a significant improvement of the model.

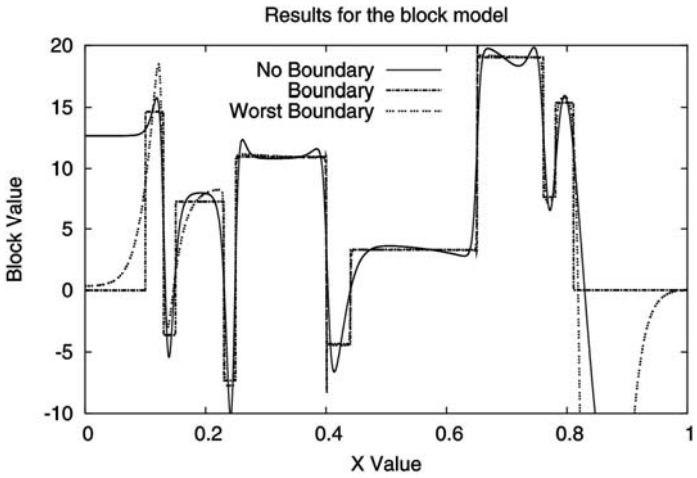

Figure 3. Resulting model output using boundary and no-boundary training data

3.0 AN ECOLOGICAL EXAMPLE

This section will describe a simple time-series model for predicting chlorophyll-*a* abundance in a freshwater lake. Boundary conditions will be defined over some of the water quality measures that can be easily related to values of chlorophyll-*a*, with the expectation that improvements in model performance should be achieved.

Lake Kasumigaura is situated in the South-Eastern part of Japan. It is a large, shallow water body where no thermal stratification occurs. Water temperatures vary widely, from 4°C in the winter to 30°C in summer. The lake has high external and internal nutrient loadings and therefore primary productivity is high. A number of climatic and limnological variables have been collected over a 10 year period (1984-1993 inclusive) for Kasumigaura, as shown in Table 1. A simple linear interpolation has been used to fill missing values to produce a complete daily time series for this period. For the purposes of this example, the years 1986 and 1993 have been kept as test data, with the remaining 8 years used for training the neural network. The aim of the neural network model was to predict chlorophyll-*a* based on inputs of Ortho phosphorus, nitrate, water temperature, Ph, dissolved oxygen and secci depth (a measure of turbidity). The model used daily data and did not incorporate any temporal features (i.e. using values from the past to support the current prediction). Hence the model produced a prediction of current chlorophyll-*a* given a set of water quality measures for a particular time instance.

Variable	Average	Units
Ortho Phosphate	14.14 ± 25.71	mg/l
Nitrate	520.56 ± 503.4	mg/l
Secchi Depth	85.43 ± 44.57	cm
Dissolved Oxygen	11.2 ± 2.14	mg/l
pH	8.74 ± 0.59	-
Water Temperature	16.36 ± 7.79	°C
Chlorophyll-*a*	74.43 ± 42.51	ug/l

Table 1. Factors measured with the daily time series data.

The parameters used with Neuroet are shown in Figure 4. Note that 20% of the original training data was used to determine when the neural network could halt training before overfitting. For both the non-boundary and boundary test cases 30 models were created.

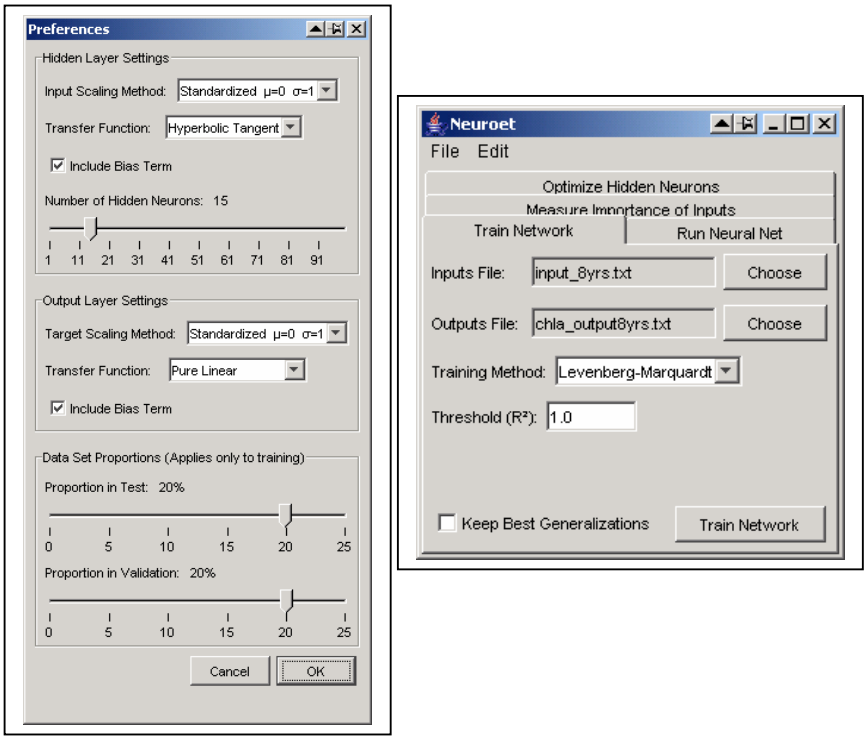

Figure 4. Neuroet parameters for all ecological model predictions

The original model, using the 8 years of training data (no boundary conditions), produced an R^2 value for the two years of test data of 0.378 ± 0.048, with a best model of 0.44.

3.1 Boundary conditions for ecological data

Ecological data is an ideal domain for defining boundary conditions. This paper will demonstrate the approach for two simple cases, but of course there are likely to be many boundary conditions based on background knowledge of the system being modeled.

The simple conditions that were selected related to Ph and dissolved oxygen:

1. if Ph is zero, then chlorophyll-*a* is 0 (highly acidic environment)
2. if Ph is 14, then chlorophyll-*a* is 0 (highly alkaline environment)
3. if Dissolved Oxygen is 0, then chlorophyll-*a* is 0 (no support for photosynthesis)

For each of these conditions, 100 training examples were produced. Each of these training examples selected random values from instances of the training data to set the other parameter values that were required as input to the model.

The resulting average R^2 value over 30 runs was 0.423 ± 0.075, with a best model of 0.56. This is a significant improvement over the non-boundary model at the 95% level for a double-sided t-test. Figure 5 shows the best runs for both the boundary and non-boundary models against the two years of test data. There is clearly a difference in the behaviour of the models, and although the peak chlorophyll-*a* value is more closely modeled with the non-boundary model, the boundary model tracks the overall patterns of the concentration more closely. Of course this example has only used 3 conditions, and the number of additional training examples that were used to represent these boundary values was selected with little knowledge of the appropriate number to best influence the training of the neural network. These issues will be discussed in §5. Figure 6 shows the R^2 values for all 30 runs of the neural network. The main point to note here is that 10 of the boundary condition runs were above the best non-boundary model. The line drawn across the graph, just above the best non-boundary model result, shows this differentiation clearly.

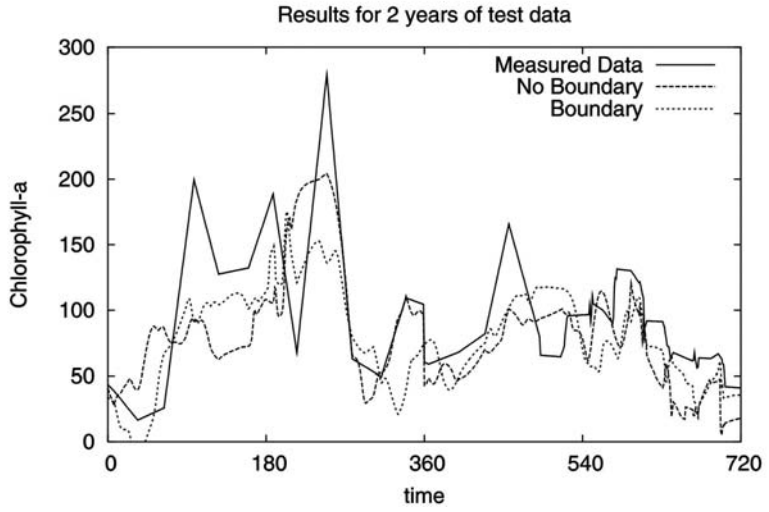

Figure5. Best models from the boundary and non-boundary training sets

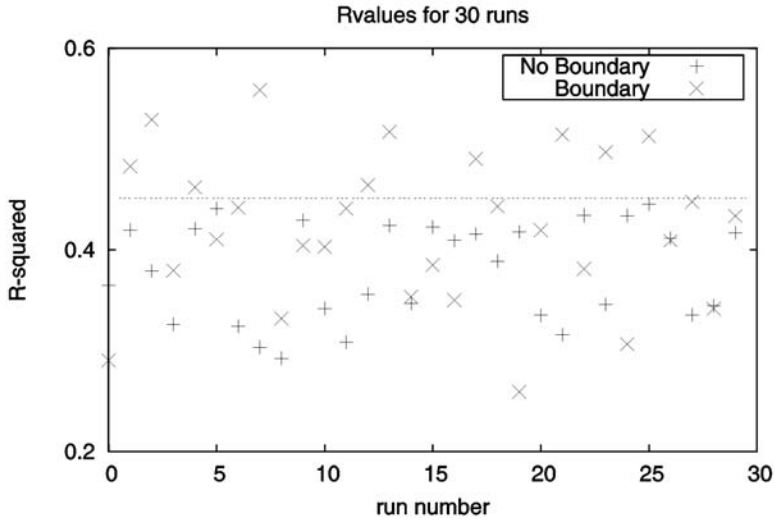

Figure 6. R^2 values for 30 model instantiations

4.0 APPLICATIONS IN SPACE AND SPACE-TIME MODELLING

The introduction of spatial concepts to a model allows a variety of boundary conditions to be potentially stated. For example, as distance becomes very large or very small, it is often possible to state some conditions of the dependent variable(s). Alternatively, topological relationships may be able to be used to state some conditions that are either never possible, or will result in known values of the independent variable(s). For example, previous work on predicting the spatial habitat preferences of marsh-breeding birds (Ozesmi & Mitsch 1997; Ozesmi & Ozesmi 1999) used a set of variables including the spatial properties of distance to open water and distance to the edge of the environment. As the distance to open water increased beyond a certain limit it was known that there was little probability of a suitable breeding site, which could easily have been incorporated as a boundary condition. In addition, since the birds preferred to breed within the habitat (edge avoiding), a second boundary condition would be that when the distance to the edge decreases to zero the breeding site probability would reduce to zero.

Another possible area where boundary conditions may be specified is in the use of remote sensed images for classification. The extreme values for some bands are likely to be able to be determined as being either unlikely

or inappropriate for the modeling problem. Specifying a set of boundary conditions that show which band values cannot occur singularly or together, in relation to dependent variable values, may allow a more generalized model to be created.

As a final example, Kanevski et al. conducted research which aimed to produce a high quality map of the variation in radioactive soil contamination caused by the Chernobyl fallout (Kanevski, Parkin, Pozdnukhov et al. 2002). Simple spatial boundary conditions could be stated for this problem, such as fallout levels decrease to zero as the distance increases beyond a certain value, and that the variogram also levels out at a certain distance (assuming that variogram values of measured data are used as input to the model).

5.0 DISCUSSION AND CONCLUSION

This paper has argued that the use of boundary conditions may improve the generalization behaviour of models constructed by supervised learning algorithms. Although this is a brief introductory study to this concept, the results seem favourable and should be explored in more detail. The main issue that needs to be addressed relates to how the boundary conditions are expressed. In particular, the number of cases for each boundary condition has to be determined so that the resulting model is not overly biased by the examples. This may be a simple rule of thumb based on a percentage of the original training data, however this is likely to be different depending on the type of learning algorithm employed. A second issue relates to the value of the other independent variables when boundary conditions are defined. Should they be randomly drawn from the training data, or as a Gaussian distribution of each variable based on the mean and standard deviation of the training data, or some other scheme? Clearly an empirical study is required for a variety of model types and data sets to determine the appropriate use of boundary conditions. A second issue is to explore the conditions under which boundary conditions may be unsuitable. Finally, a study is required for a variety of spatial and temporal domains so that guidelines can be produced indicating the types of boundary conditions that are likely to be useful when developing inductive models.

REFERENCES

Hirsh, H. & M. Noordewier (1994) Using Background Knowledge to Improve Inductive Learning. *IEEE Intelligent Systems*, 9:5, pp. 3-6.

Kanevski, M., R. Parkin, A. Pozdnukhov, V. Timonin, M. Maignan, B. Yatsalo & S. Sanu (2002) Environmental Data Mining and Modelling Based on Machine Learning Algorithms and Geostatistics. *Integrated Assessment and Decision Support: Proceedings of the 1st Biennial Meeting of the iEMSs*, .

Langley, P. (1986) On Machine Learning. *Machine Learning*, 1:1, pp. 5-10.

Lenat, D. 1984, 'The Role of Heuristics in Learning by Discovery: Three Case Studies', in *Machine Learning: An Artificial Intelligence Approach*, Ed R. S. a. C. Michalski, J.G. and Mitchell,T.M, pp. 243-306.

Muggleton, S. & W. Buntine 1992, 'Machine Invention of First-Order Predicates by Inverting Resolution', in *Inductive Logic Programming*, Ed S. Muggleton, pp. 261-281.

Ozesmi, S. & U. Ozesmi (1999) An artificial neural network approach to spatial habitat modelling with interspecific interaction. *Ecological Modelling*, 116: pp. 15-31.

Ozesmi, U. & W. Mitsch (1997) A spatial habitat model for the marsh-breeding red-winged blackbird (Agelaius phoeniceus L.) in coastal Lake Erie wetlands. *Ecological Modelling*, 101: pp. 139-152.

Rendell, L. & H. Cho (1990) Empirical learning as a function of concept character. *Machine Learning*, 5:3, pp. 267-298.

Sarle, W. S. 1999, *Donoho-Johnstone benchmarks: neural net results*, fpt://ftp.sas.com/pub/neural/dojo/dojo.html, Last accessed 1/10/2004.

Silvert, W. & M. Baptist 1998, 'Can Neural Networks be used in Data-Poor Situations?' in *Artificial Neuronal Networks: Application to Ecology and Evolution*, Eds S. Lek & J. Guegan, Springer-Verlag, Berlin, pp. 241-248.

Tribou, E. & P. Nobel 2004, *Neuroet: a simple artificial neural network for scientists*, Civil and Environmental Engineering, University of Washington, City, pp 43.

Utgoff, P. 1986, 'Shift of bias for inductive concept learning', in *Machine Learning: An Artificial Intelligence Approach* Morgan Kaufmann Publishers, Inc. Los Altos, CA, pp. 107-148.