Journal of Neuroscience Methods 220 (2013) 85–91

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

journal homepage: www.elsevier.com/locate/jneumeth

Basic Neuroscience

A comparison of random forest regression and multiple linear regression for prediction in neuroscience

Paul F. Smith[a,*], Siva Ganesh[c], Ping Liu[b]

[a] Department of Pharmacology and Toxicology, The Brain Health Research Centre, University of Otago, Dunedin, New Zealand
[b] Anatomy, School of Medical Sciences, The Brain Health Research Centre, University of Otago, Dunedin, New Zealand
[c] Bioinformatics and Statistics, AgResearch Ltd., Palmerston North, New Zealand

HIGHLIGHTS

- Multiple linear regression is often used for prediction in neuroscience.
- Random forest regression is an alternative form of regression.
- It does not make the assumptions of linear regression.
- We show that linear regression can be superior to random forest regression.

ARTICLE INFO

Article history:
Received 22 May 2013
Received in revised form 13 August 2013
Accepted 28 August 2013

Keywords:
Regression
Linear regression
Regression trees
Random forest regression
L-Arginine metabolism
Vestibular nucleus
Cerebellum

ABSTRACT

Background: Regression is a common statistical tool for prediction in neuroscience. However, linear regression is by far the most common form of regression used, with regression trees receiving comparatively little attention.
New method: In this study, the results of conventional multiple linear regression (MLR) were compared with those of random forest regression (RFR), in the prediction of the concentrations of 9 neurochemicals in the vestibular nucleus complex and cerebellum that are part of the L-arginine biochemical pathway (agmatine, putrescine, spermidine, spermine, L-arginine, L-ornithine, L-citrulline, glutamate and γ-aminobutyric acid (GABA)).
Results: The R^2 values for the MLRs were higher than the proportion of variance explained values for the RFRs: 6/9 of them were ≥0.70 compared to 4/9 for RFRs. Even the variables that had the lowest R^2 values for the MLRs, e.g. ornithine (0.50) and glutamate (0.61), had much lower proportion of variance explained values for the RFRs (0.27 and 0.49, respectively). The RSE values for the MLRs were lower than those for the RFRs in all but two cases.
Comparison with existing methods: In general, MLRs seemed to be superior to the RFRs in terms of predictive value and error.
Conclusion: In the case of this data set, MLR appeared to be superior to RFR in terms of its explanatory value and error. This result suggests that MLR may have advantages over RFR for prediction in neuroscience with this kind of data set, but that RFR can still have good predictive value in some cases.

1. Introduction

Linear regression is a part of the general linear model (GLM) that is often used to predict one variable from another in neuroscience. Simple linear regression can be expanded to include more than one predictor variable to become multiple linear regression. However, formal statistical tests of multiple linear regression, like simple linear regression, make assumptions regarding the distribution of the data, which cannot always be fulfilled. These assumptions are that the data are normally distributed, with homogeneity of variance, and that they are independent of one another (e.g. not autocorrelated) (Vittinghoff et al., 2005). Furthermore, the predictor variables should be numerical, although indicator variables can be used in order to include nominal variables (e.g., binary coding to represent male and female). The violation of the assumption of normality can sometimes be redressed using data transformation, which may also correct heterogeneity of variance, but other issues such as autocorrelation are not easily dealt with and may require methods such as time series regression (Ryan, 2009).

Although modelling using regression trees has been used for over 25 years, its use in the neurosciences has been very limited. In regression tree modelling, a flow-like series of questions is asked

* Corresponding author. Tel.: +64 3 479 5747.
E-mail address: paul.smith@stonebow.otago.ac.nz (P.F. Smith).

0165-0270/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jneumeth.2013.08.024

about each variable ('recursive partitioning'), subdividing a sample into groups that are as homogeneous as possible by minimising the within-group variance, in order to determine a numerical response variable (Vittinghoff et al., 2005). The predictor variables can be numerical also, or they can be ordinal or nominal. By contrast with linear regression, no assumptions are made about the distribution of the data. The data are usually split into training and test data sets (e.g., 90:10) and the mean square error (MSE) between the model based on the training data and the test data is calculated as a measure of the model's success. Variables are chosen to split the data based on the reduction in the MSE achieved after a split (i.e., the information gained). Unlike linear regression, interactions between different predictor variables are automatically incorporated into the regression tree model and variable selection is unnecessary because irrelevant predictors are excluded from the model. This makes complex, non-linear interactions between variables easier to accommodate than in linear regression modelling (Hastie et al., 2009). Breiman et al. (1984) extended the concept of regression trees by exploiting the power of computers to simultaneously generate hundreds of regression trees, known as 'random forests', which were based on a random selection of a subset of data from the training set. The various regression tree solutions are averaged in order to predict the target variable with the smallest MSE (Marsland, 2009).

The aim of this study was to compare the results of a conventional multiple linear regression with those of random forest regression, using data on the expression of neurochemicals related to the L-arginine metabolic pathway in the rat hindbrain as an example. Two areas of the hindbrain concerned with the control of movement were investigated: the brainstem vestibular nucleus complex (VNC) and the cerebellum (CE), in young (4 month old) and aged (24 month old) rats (Liu et al., 2010). Chemical analyses were performed to determine the concentrations of 9 related neurochemicals that form a biochemical pathway that is critical for neuronal function (see Fig. 1): agmatine, putrescine, spermidine, spermine, L-arginine, L-ornithine, L-citrulline, glutamate and γ-aminobutyric acid (GABA). Although Fig. 1 presents certain causal connections between some of these neurochemical variables, the mechanisms through which they interact with one another are not completely understood and additional pathways, particularly feedback pathways, are possible (Mori and Gotoh, 2004). It is therefore of interest to determine whether the concentrations of one part of this complex neurochemical pathway can be predicted from the other parts.

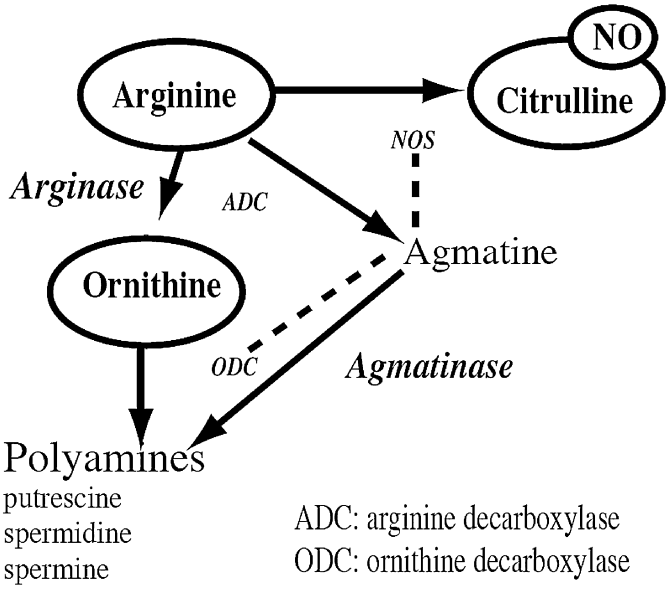

Fig. 1. The arginine metabolic pathway showing the conversion of L-arginine to the neurotransmitter, nitric oxide (NO), and L-citrulline, by the enzyme, nitric oxide synthase (NOS), of which there are 3 isoforms; the conversion of L-arginine to agmatine by the enzyme, arginine decarboxylase (ADC), which is then converted to polyamines such as putrescine, spermidine and spermine by agmatinase and ornithine decarboxylase (ODC); and the conversion of L-arginine to L-ornithine by arginase, which is then converted to the same polyamines, which are essential for cell proliferation, differentiation and communication, including neuronal synaptic plasticity in the brain. The major excitatory neurotransmitter, glutamate, is one of the end products of L-arginine, and glutamate serves as a precursor for the synthesis of the major inhibitory neurotransmitter, GABA. Therefore, all of these neurochemicals are interconnected.

2. Methods

2.1. Data set and variables

The data set was obtained from Liu et al. (2010). Male Sprague-Dawley rats (aged: 24 months old, $n = 14$; young: 4 months old, $n = 14$) were housed 3–5 per cage and maintained on a 12 h light-dark cycle and provided with ad lib. access to food and water. All experimental procedures were carried out in accordance with the regulations of the University of Otago Committee on Ethics in the Care and Use of Laboratory Animals. Animals were housed either in a standard rat cage or an enriched environment including toys and other novel objects, since enriched environments have been shown to reduce age-related memory impairment (Olson et al., 2006). Therefore, the sample sizes for the aged and young groups were divided according to the housing conditions. In order to achieve as large a sample size as possible, data from the VNC and CE were combined in the regression analyses, so that for each of the 9 neurochemical variables the total n was 58. This was considered to be a reasonable solution given the close physiological relationship between the VNC and CE (Liu et al., 2010). This meant that for the aged group with standard housing, n was = 13, aged with enriched housing, n was = 16; young with standard housing, n was = 14, and for young with enriched housing, n was = 15. These smaller sample sizes were less important because age and enrichment were categorical variables that were never the target variables, but they were included in the regression analyses as predictor variables. A previous study using the same data set analysed the data using multivariate analyses of variance (MANOVAs), linear discriminant and cluster analyses (Liu et al., 2010), but the main interest in the latter case was the prediction of the age of the brain tissue based on the other variables rather than predicting neurochemical concentrations using regression analyses. Determination of the concentrations of agmatine, putrescine, spermidine, spermine, L-arginine, L-ornithine, L-citrulline, glutamate and γ-aminobutyric acid (GABA) was carried out using high performance liquid chromatography (HPLC) or a highly sensitive liquid chromatography/mass spectrometric (LC/MS/MS) method and expressed as μg/g of wet tissue weight (see Liu et al., 2008a, 2010 for details).

The experimental design thus consisted of 2 main independent variables: age with 2 levels, 4 months old and 24 months old; and housing, with 2 levels, standard and enriched. There were 9 potential dependent variables corresponding to the concentrations of agmatine, putrescine, spermidine, spermine, L-arginine, L-ornithine, L-citrulline, glutamate and GABA. However, in any one regression analysis, only one of these continuous neurochemical variables was the target or y variable and the other 8 were included as predictor variables. Consequently, each analysis involved 10 predictor variables, i.e. 8 continuous variables and 2 categorical ones, and one dependent continuous neurochemical variable.