Jurnal Natur Indonesia 11(1), Oktober 2008: 64-69
ISSN 1410-9379, Keputusan Akreditasi No 55/DIKTI/Kep./2005

Statistical Significance Test for Neural Network Classification

Sri Rezeki[1*], Subanar[2] dan Suryo Guritno[2***]**

[1] Department of Mathematics Education, Universitas Islam Riau, Pekanbaru, Indonesia
[2] Department of Mathematics, Universitas Gadjah Mada, Yogyakarta, Indonesia

Diterima 04-10-2007 Disetujui 09-10-2008

ABSTRACT

Model selection in neural networks can be guided by statistical procedures, such as hypothesis tests, information criteria and cross validation. Taking a statistical perspective is especially important for nonparametric models like neural networks, because the reason for applying them is the lack of knowledge about an adequate functional form. Many researchers have developed model selection strategies for neural networks which are based on statistical concepts. In this paper, we focused on the model evaluation by implementing statistical significance test. We used Wald-test to evaluate the relevance of parameters in the networks for classification problem. Parameters with no significance influence on any of the network outputs have to be removed. In general, the results show that Wald-test work properly to determine significance of each weight from the selected model. An empirical study by using Iris data yields all parameters in the network are significance, except bias at the first output neuron.

Keywords: parameter significance, Wald-test, classification.

INTRODUCTION

One of the most unresolved questions in the literature on neural networks (NN) is what architecture should be used for a given problem. Architecture selection requires choosing both the appropriate number of hidden units and the connection thereof (Sarle 1994). A desirable network architecture contains as few hidden units and connections as necessary for a good approximation of the true function, taking into account the trade-off between estimation bias and variability due to estimation errors. It is therefore necessary to develop a methodology to select an appropriate network model for a given problem.

Reed (Reed 1993) provides a survey about the usual approaches pursued in the network literature. The approaches, for example, are regularization, pruning, and stopped training. Regularization methods choose the network weights such that they minimize the network error function (e.g. sum of squared errors) plus a penalty term for the networks complexity. Another way to justify the regularization term is to formalize and interpret the method in a Bayesian framework. This was reviewed for example in Bishop (1995) and Ripley (1996).

*Telp: 081276430115
Email: *sri_rezeki_uir@yahoo.com, **subanar@yahoo.com, ***guritno0@mailcity.ac.id

Pruning methods identify the parameters that do not 'significantly' contribute to the overall network performance. However, this 'significance' is usually not judged on the basis of test statistics (an exception is Burgess (1995), who uses F-tests to identify insignificant parts of a network model). Pruning methods use the so-called 'saliency' as a measure of a weight's importance. The saliency of a weight is defined as the increase in network model error (e.g. sum of squared errors) incurred by setting this weight to zero. The idea is to remove the weights with low saliency; however, the method does not provide any guidelines as to whether or not a saliency should be judged as low. It is shown that the computation of the saliency is generalized by a corresponding Wald-test statistics (Anders & Korn 1996). An alternative to conventional pruning methods is developed and tested in Kingdon (1997). The basic idea of the approach, called 'network regression pruning', is to remove network weights while retaining the network's mapping. A weight is seen as redundant if, after its removal, the original mapping can be (approximately) recovered by adjusting the remaining weights of the effected node.

In the application of stopped training, the data set is split into a training and a validation set. If the model errors in the validation set begin to grow during the training process, the training algorithm is stopped. In statistical terms, the method tries to make up the

R^v, $l(z,\cdot)$ is continuous on *W*. Suppose further that there exists $d:\mathrm{R}^3 \to \mathrm{R}^+$ such that for all *w* in *W* $|l(z,w)| \le d(z)$ and *E*(d(Zt)) < ∞ (i.e., *l* is dominated on *W* by an integrable function).

Then for each *n* = 1, 2, ... there exists a solution $\hat{w}_n$ to the problem

$$\min_{w\in W} \hat{a}_n(w) = n^{-1}\sum_{t=1}^{n} l(Z_t, w) \text{ and } \hat{w}_n \xrightarrow{a.s} w^*,$$

where:

$$w^* \equiv \{w^* \in W : a(w^*) \le a(w)\},\ a(w) \equiv E(l(Z_t, w))$$

Theorem 2.2. (White 1989). Let (Ω, F, P), $\{Z_t\}$, W and l be as in Theorem 2.1, and suppose that $\hat{w}_n \xrightarrow{a.s} w^*$ where w^* is an isolated element of W^* interior to W. Suppose in addition that for each z in R^v, $l(z,\cdot)$ is continuously differentiable of order 2 on $\int W$; that $E(\nabla l(Z_t, w^*)'\nabla l(Z_t, w^*)) < \infty$; that each element of $\nabla^2 l$ is dominated on W by an integrable function; and that $\mathbf{A}^* \equiv E(\nabla^2 l(Z_t, w^*))$ and $\mathbf{B}^* \equiv E(\nabla l(Z_t, w^*)\nabla l(Z_t, w^*)')$ are nonsingular $(s \times s)$ matrices, where ∇ and ∇^2 denote the $(s \times 1)$ gradient and $(s \times s)$ Hessian operators with respect to w.

Then $\sqrt{n}(\hat{w}_n - w^*) \xrightarrow{d} N(0, \mathbf{C}^*)$, where $\mathbf{C}^* = \mathbf{A}^{*-1}\mathbf{B}^*\mathbf{A}^{*-1}$. If in addition each element of $\nabla l \nabla l'$ is dominated on W by an integrable function, then $\hat{\mathbf{C}}_n \xrightarrow{a.s.} \mathbf{C}^*$, where $\hat{\mathbf{C}}_n = \hat{\mathbf{A}}_n^{-1}\hat{\mathbf{B}}_n\hat{\mathbf{A}}_n^{-1}$, and

$$\hat{\mathbf{A}}_n = \frac{\sum_{t=1}^{n} \nabla^2 l(Z_t, \hat{w}_n)}{n},$$

$$\hat{\mathbf{B}}_n = \frac{\sum_{t=1}^{n} \nabla l(Z_t, \hat{w}_n)\nabla l(Z_t, \hat{w}_n)'}{n}.$$

Corollary 2.3. (White 1999) Let $\{\mathbf{b}_n\}$ be a sequence of random $k \times 1$ vectors such that $\mathbf{V}_n^{-1/2}\mathbf{b}_n \xrightarrow{d} N(0, \mathbf{I})$, where $\{\mathbf{V}_n\}$ and $\{\mathbf{V}_n^{-1}\}$ are $O(1)$. Let $\{\mathbf{A}_n\}$ be a $O(1)$ sequence of (nonstochastic) $q \times k$ matrices with full row rank *q* for all n sufficiently large, uniformly in n. Then the sequence $\{\mathbf{A}_n\mathbf{b}_n\}$ is such that

$$\mathbf{\Gamma}_n^{-1/2}\mathbf{A}_n\mathbf{b}_n \xrightarrow{d} N(0, \mathbf{I}),$$

where $\mathbf{\Gamma}_n \equiv \mathbf{A}_n\mathbf{V}_n\mathbf{A}_n'$ and $\mathbf{\Gamma}_n$ and $\mathbf{\Gamma}_n^{-1}$ are $O(1)$.

Proposition 2.4. (White 1999) Let $g: R^k \to R^l$ be continuous on a compact set $C \subset R^k$. Suppose that $\{\mathbf{b}_n\}$ is a sequence of random vector $k \times 1$ and $\{\mathbf{c}_n\}$ is a sequence of $k \times 1$ vector such that $\mathbf{b}_n - \mathbf{c}_n \xrightarrow{p} 0$, and for all n sufficiently large, $\mathbf{c}_n$ is interior to $\mathbf{C}$, uniformly in n.

Then $g(\mathbf{b}_n) - g(\mathbf{c}_n) \xrightarrow{p} 0$.

Theorem 2.5. (White 1999) Let $\mathbf{V}_n^{-1/2}\mathbf{b}_n \xrightarrow{d} N(0, \mathbf{I}_k)$, and suppose there exists $\hat{\mathbf{V}}_n$ positive semi definite and symmetric such that $\hat{\mathbf{V}}_n - \mathbf{V}_n \xrightarrow{p} 0$, where $\mathbf{V}_n$ is $O(1)$, and for all n sufficiently large, $\det(\mathbf{V}_n) > ¢ > 0$. Then $\mathbf{b}_n'\hat{\mathbf{V}}_n'\mathbf{b}_n \xrightarrow{d} ¡_k^2$.

The Wald statistic allows the simplest analysis, although it may or may not the easiest statistic to compute in a given situation. The motivation for the Wald statistic is that when the null hypothesis is correct $\mathbf{S}\hat{\mathbf{w}}_n$ should be close to $\mathbf{S}\mathbf{w}^* = \mathbf{s}$, so a value of $\mathbf{S}\hat{\mathbf{w}}_n - \mathbf{s}$ far from zero is evidence against the null hypothesis.

The theorem about Wald statistic that be used for hypothesis testing of parameters in NN model is adapted from White (1999) Theorem 4.31. The result of adaptation for our specific purpose is as follow:

Theorem 2.6. Let the conditions of Theorem 2.2 hold, i.e.:

(i) $\mathbf{C}^{*-1/2}\sqrt{n}(\hat{\mathbf{w}}_n - \mathbf{w}^*) \xrightarrow{d} N(\mathbf{0}, \mathbf{I})$, where $\mathbf{C}^* \equiv \mathbf{A}^{*-1}\mathbf{B}^*\mathbf{A}^{*-1'}$, and $\mathbf{C}^{*-1}$ is $O(1)$,

(ii) there exists a matrix $\hat{\mathbf{B}}_n$ positive semi definite and symmetric such that $\hat{\mathbf{B}}_n - \mathbf{B}^* \xrightarrow{p} \mathbf{0}$.

Then $\hat{\mathbf{C}}_n - \mathbf{C}^* \xrightarrow{p} \mathbf{0}$, where

$\hat{\mathbf{C}}_n = \hat{\mathbf{A}}_n^{-1}\hat{\mathbf{B}}_n\hat{\mathbf{A}}_n^{-1}$, and

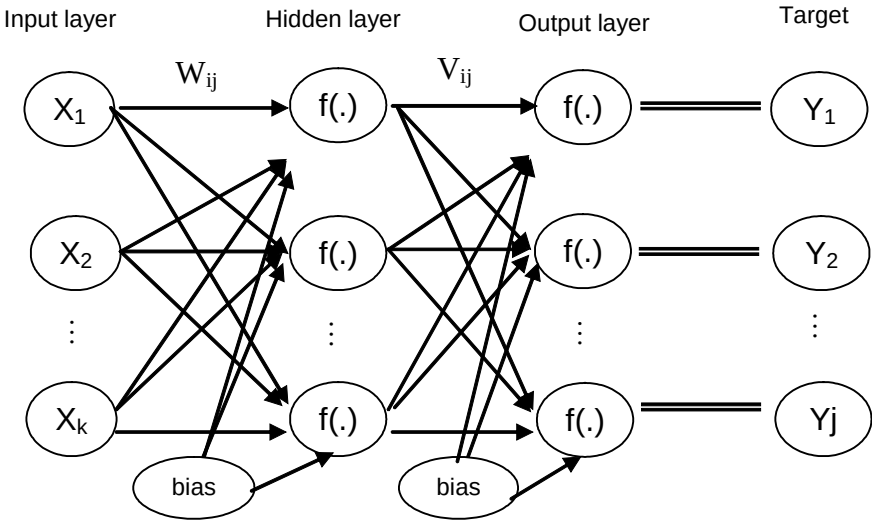

Figure 1a. FFNN with single hidden layer

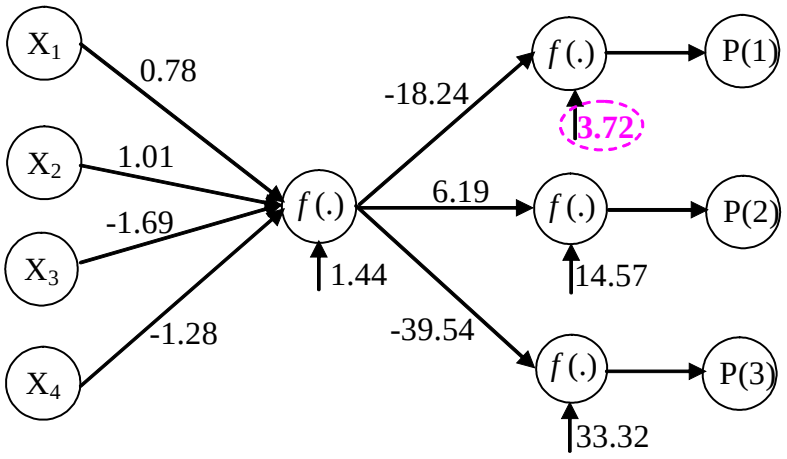

Figure 1b. Architecture of NN (4-1-3) with the value of weights estimation

Table 1. Wald test of weights estimation for NN (4-1-3)

Weight	Coefficient	S.E.	Wald test	p-value
b→h1	1.44	0.33496	18.480	0.00002
i1→h1	0.78	0.07969	95.766	0.00000
i2→h1	1.01	0.08373	145.570	0.00000
i3→h1	-1.69	0.12108	194.825	0.00000
i4→h1	-1.28	0.13164	94.544	0.00000
b→o1	3.72	2.51477	2.188	0.13907
h1→o1	-18.24	2.76089	43.647	0.00000
b→o2	14.57	2.74225	28.230	0.00000
h1→o2	6.19	2.78325	4.946	0.02615
b→o3	33.32	2.60121	164.081	0.00000
h1→o3	-39.54	3.09831	162.864	0.00000

significance of these estimators are presented at Table 1.

Usually, neurons in NN are always full connectivity. By using this Wald test, it is possible that there are weights which are not significance. Based on Table 1, all of the weights are significance except bias at the first output neuron. Therefore, the connection should be removed and the best model is not fully connected.

CONCLUSION

In general, neural networks are applied to problems where slight is known about the correct functional form. Therefore, a statistical approach to the model selection seems particularly important and should become an integral part of neural networks modeling. Wald-test work properly in theoretically because $\hat{w}$ is asymptotically normal with mean w^* and covariance matrix $(1/n)\mathbf{C}^*$, or $\sqrt{n}(\hat{w}-w^*) \sim N(0,\mathbf{C}^*)$. Empirically, Wald-test is also applicable for determining parameter significance of the selected neural networks