Publication VI

Emil Eirola, Elia Liitiäinen, Amaury Lendasse, Francesco Corona, and Michel Verleysen. Using the Delta test for variable selection. In *European Symposium on Artificial Neural Networks (ESANN 2008)*, pages 25–30, April 2008.

ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

Using the Delta Test for Variable Selection

Emil Eirola[1], Elia Liitiäinen[1], Amaury Lendasse[1],
Francesco Corona[1], and Michel Verleysen[2]

1—Helsinki University of Technology—Lab. of Computer and Information Science
P.O. Box 5400, FI–2015 HUT—Espoo, Finland.

2—Université Catholique de Louvain—Machine Learning Group
Place du Levant 3, B–1348—Louvain-la-Neuve, Belgium

Abstract. Input selection is an important consideration in all large-scale modelling problems. We propose that using an established noise variance estimator known as the Delta test as the target to minimise can provide an effective input selection methodology. Theoretical justifications and experimental results are presented.

1 Introduction

In regression analysis, two of the main concerns are accuracy of the model and increased interpretability of the data. Input selection is one way to address both of these issues. The constantly increasing size of relevant data sets, however, requires progressively more sophisticated methods for input selection. Several popular methods exist for this task (e.g., [1]), but many suffer from the "curse of dimensionality" in some way or another when the data sets escalate to very high dimensionalities [2]. We investigate a method based on the concept of *nearest neighbors* (NN) to evaluate input selections, since intuitively such proximity-measures are less affected by this curse. The effects of dimensionality related to NNs have been extensively studied in the literature [3, 4].

Our method is based on a well-known noise variance estimator commonly referred to as the Delta test [5, 6]. Intuitively, it seems sensible to compare different input selections by minimising a noise estimate, and this method has been used with some success [7]. In this paper, we present some justification for this procedure, and explanations for why in particular the Delta test is an appropriate estimator to use for this task, when it is well known that there are more sophisticated methods for actual noise estimation—e.g., [8, 9].

The goal of this paper is to provide a mathematically convincing—though not technically rigorous—argument supported by an example as to why the suggested methodology is valid.

This paper is organised as follows: in Section 2, we review some fundamentals of noise variance estimation. Section 3 comprises the main result, a theoretical analysis of why the Delta test can be used for variable selection. In Section 4, we present some supporting experimental evidence.

2 Noise Variance Estimation

In function approximation, we have a set of input points $(x_i)_{i=1}^M$ and associated scalar outputs $(y_i)_{i=1}^M$. The assumption is that there is a functional dependence

between them, but with an additive noise term:

$$y_i = f(x_i) + \varepsilon_i$$

The function f is assumed to be smooth, and the noise terms ε_i are i.i.d. with zero mean. Noise variance estimation is the study of how to give an *a priori* estimate for $\mathrm{Var}(\varepsilon)$ given some data *without* considering any specifics of the shape of f.

2.1 Nearest Neighbors

The noise variance estimator considered here is based on a nearest neighbour (NN) approach. The NN of a point is defined as the (unique) point which minimises a distance metric to that point *in the input space*:

$$N(i) := \arg\min_{j \neq i} \|x_i - x_j\|^2$$

In this context, we use the Euclidean distance, but other metrics can also be used.

2.2 The Delta Test

The Delta test [5, 6] is usually written as

$$\mathrm{Var}(\varepsilon) \approx \frac{1}{2M} \sum_{i=1}^{M} \left(y_i - y_{N(i)}\right)^2,$$

i.e., we consider the differences in the outputs associated with neighboring (in the input space) points. This is a well-known and widely used estimator, and it has been shown—e.g., in [10]—that the estimate converges to the true value of the noise variance in the limit $M \to \infty$.

2.3 Noise Variance Estimators for Input Selection?

Noise variance estimators have been used previously for variable selection procedures, but there are some essential problems with this usage that need to be addressed. Indeed, any reasonable noise estimator would manage to include all of the relevant variables, since excluding them would cause unexplainable variations in the data. However, most estimators fail at the equally important task of pruning irrelevant inputs, since in that context every variable has the possibility of containing a slight bit of additional information and including it might lead to a lower estimate.

For eliminating variables, it is then somewhat counter-intuitive to be using a noise-estimation scheme. In spite of this, we will see that the Delta test has the interesting property that adding unrelated inputs *does* increase the estimate, separating it from most other estimators, and making it effective for input selection.

ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

3 Theoretical Justification

In this section, we present the rationale for why minimising the Delta test can be an effective variable selection procedure. Let the input space dimension $d \in \mathbb{N}$ and the number of points $M \in \mathbb{N}$ be fixed.

We first assume that our data $x_i \in [0,1]^d$ for $1 \leq i \leq M$ are i.i.d. uniformly distributed on the unit hypercube. Consequently the components (or variables, denoted x_i^k) of each x_i are i.i.d. on the interval $[0,1]$. Let $y_i := f(x_i) + \varepsilon_i$ for $1 \leq i \leq M$ where $f : [0,1]^d \to \mathbb{R}$ is a continuous function with bounded first and second partial derivatives. The residuals ε_i are i.i.d. random variables with zero mean and $\mathrm{Var}(\varepsilon_i) = \sigma^2$. The points $(x_i)_{i=1}^M$ and $(y_i)_{i=1}^M$ now comprise our imitation data set.

In general, there will be some inputs for f which are not significant—denote by $D \in \mathcal{P}(\{1,...,d\})$ the set of variables which truly affect the output:

$$D = \{k \mid \partial_k f \text{ is non-zero somewhere}\}$$

Define the Delta test $\delta : \mathcal{P}(\{1,...,d\}) \to \mathbb{R}$ so that

$$\delta(S) := \frac{1}{2M} \sum_{i=1}^{M} \left(y_i - y_{N_S(i)}\right)^2$$

where

$$N_S(i) := \arg\min_{j \neq i} \|x_i - x_j\|_S^2 ,$$

and the seminorm

$$\|x_i - x_j\|_S^2 := \sum_{k \in S} \left(x_i^k - x_j^k\right)^2 .$$

This representation of the Delta test maps each selection $S \subset \{1,...,d\}$ of variables to an estimate for the noise where the nearest neighbour is calculated in the subspace spanned by the variables specified in S.

Conjecture 1 *The correct selection of variables uniquely minimises the expected value of the Delta test.*

$$S \neq D \implies \mathrm{E}\left[\delta(S)\right] > \mathrm{E}\left[\delta(D)\right]$$

Sketch of proof.

$$\begin{aligned}
\mathrm{E}\left[\delta(S)\right] =& \mathrm{E}\left[\frac{1}{2M} \sum_{i=1}^{M} \left(y_i - y_{N_S(i)}\right)^2\right] = \frac{1}{2}\mathrm{E}\left[\left(y_i - y_{N_S(i)}\right)^2\right] \\
=& \frac{1}{2}\mathrm{E}\left[\left(f(x_i) - f(x_{N_S(i)}) + \varepsilon_i - \varepsilon_{N_S(i)}\right)^2\right] \\
=& \frac{1}{2}\mathrm{E}\left[\left(f(x_i) - f(x_{N_S(i)})\right)^2\right] + \sigma^2,
\end{aligned}$$

since the ε terms are independent from x and each other. It then suffices to show that

$$S \neq D \implies \mathrm{E}\left[\left(f(x_i) - f(x_{N_S(i)})\right)^2\right] > \mathrm{E}\left[\left(f(x_i) - f(x_{N_D(i)})\right)^2\right].$$

Linear f

To illustrate the idea of the proof, we first study the case of a linear f, that is, $f(x_i) = a_0 + \sum_{k \in D} a_k x_i^k$ with $a_k \neq 0$ for $k \in D$. Now

$$\mathrm{E}\left[\left(f(x_i) - f(x_{N_S(i)})\right)^2\right] = \mathrm{E}\left[\left(\sum_{k \in D} a_k \left(x_i^k - x_{N_S(i)}^k\right)\right)^2\right]$$

and since the components are uncorrelated:

$$= \mathrm{E}\left[\sum_{k \in D} a_k^2 \left(x_i^k - x_{N_S(i)}^k\right)^2\right] = \sum_{k \in D} a_k^2 \,\mathrm{E}\left[\left(x_i^k - x_{N_S(i)}^k\right)^2\right]$$

$$= \sum_{k \in D \cap S} a_k^2 \underbrace{\mathrm{E}\left[\left(x_i^k - x_{N_S(i)}^k\right)^2\right]}_{=g(\#S)} + \sum_{k \in D \setminus S} a_k^2 \underbrace{\mathrm{E}\left[\left(x_i^k - x_{N_S(i)}^k\right)^2\right]}_{=1/6}$$

Here the second term is $1/6$ because x_i^k and $x_{N_S(i)}^k$ are independent and uniformly distributed on $[0, 1]$ when $k \notin S$. The function $g(\#S)$—which measures the expected distance (squared) along one component in S from a point to its nearest neighbor in the subspace of S—however, should clearly be far less than $1/6$, as long as M is large enough so that nearest neighbors can be expected to be considerably closer than randomly chosen points.

Still, $g(\#S)$ is an increasing function of the number of variables in S. This means that the expression is minimised by the *smallest* selection which includes D, so it is minimised by $S = D$.

General f

For the general case of a smooth f, we apply the mean-value theorem to give us a point $\hat{x}_i$ on the line segment between x_i and $x_{N_S(i)}$ for which

$$\mathrm{E}\left[\left(f(x_i) - f(x_{N_S(i)})\right)^2\right] = \mathrm{E}\left[\left(\nabla f\left(\hat{x}_i\right)\left(x_i - x_{N_S(i)}\right)\right)^2\right]$$

Since the components are uncorrelated, we proceed as in the linear case.

$$= \mathrm{E}\left[\left(\sum_{k \in D} \partial_k f\left(\hat{x}_i\right)\left(x_i^k - x_{N_S(i)}^k\right)\right)^2\right] = \sum_{k \in D} \mathrm{E}\left[\left(\partial_k f\left(\hat{x}_i\right)\right)^2 \left(x_i^k - x_{N_S(i)}^k\right)^2\right]$$

$$= \sum_{k \in D \cap S} \mathrm{E}\left[\left(\partial_k f\left(\hat{x}_i\right)\right)^2 \left(x_i^k - x_{N_S(i)}^k\right)^2\right] + \sum_{k \in D \setminus S} \mathrm{E}\left[\left(\partial_k f\left(\hat{x}_i\right)\right)^2 \left(x_i^k - x_{N_S(i)}^k\right)^2\right]$$

As above, the second term here will be considerably large if $D \setminus S \neq \emptyset$, since those particular variables are not considered in the minimisation but do affect the output. Hence we need $S \supset D$ for S to minimise the expression. As for

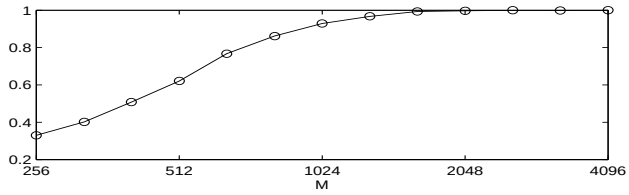

Fig. 1: The convergence of choosing the correct input selection. The vertical axis represents the ratio of cases where the Delta test correctly identified the true selection from a total of 1000 tests for each point.

the first term, the differences $x_i^k - x_{N_S(i)}^k$ will on average grow slightly with the size of S as there are more variables to take into account in the NN search. So again, the minimising selection is the smallest set which contains D, and hence the expression is uniquely minimised by $S = D$.
End of sketch of proof.

The assumptions that were made for the distribution of the data points (uniform on unit hypercube) may seem strict, and our proof is rigorous only for linear functions. However, the Delta test is a local method, and since any continuous distribution is locally uniform and any smooth function is locally linear the idea easily generalises to continuous distributions on bounded domains.

4 Experimental Results

To illustrate the effectiveness of the procedure, we conducted an artificial experiment. For this test, we intentionally used a very nonlinear function:

$$f(x_1, x_2, x_3, x_4, x_5, x_6) = \cos(2\pi x_1)\cos(4\pi x_2)\exp(x_2)\exp(2x_3)$$

for $x_i \in [0, 1]$. Obviously $D = \{1, 2, 3\}$ in this case. The variance of the noise was chosen to be 10, which is quite considerable considering the range of the data.[1] The estimator was given all $2^6 - 1$ different possible input selections, and the one which minimises the estimate is chosen. The results are presented in Figure 1, where the vertical axis represents the fraction of cases where the correct selection was chosen. The experiment was performed as a Monte Carlo simulation with 1000 repetitions for each of the different data set sizes M.

It is clear that that with increasing data size, the Delta test is eventually able to *always* choose the correct selection. The necessary size of over 1000 points in this case might seem high, but recall that the situation was deliberately chosen to be problematic. We ran the identical experiment with other noise variance estimators [8, 9], and they definitely did *not* converge to a value of 1. The reason that more sophisticated estimators can not be used for input selection in this way is precisely because they are effective at accurately estimating the noise variance—the bias of the estimate is (practically) zero even for a small number

[1] The signal-to-noise ratio $\mathrm{Var}(f(x))/\mathrm{Var}(\varepsilon)$ is approximately 1.08.

of samples. The Delta test, as shown in [10], however, *has* a slight bias which approaches zero in the limit $M \to \infty$, but for a fixed M we are able to exploit this bias for input selection purposes as explained in the previous section.

5 Conclusions

We have proposed that using the Delta test—a noise variance estimator—as the target function for an input selection procedure can give effective results. In addition to the theoretical treatment in Section 3, we provided illustrative experimental evidence.

Now that we have elementary confirmation that the procedure is valid, the next step will be to investigate further properties of the strategy. More experiments with real-world data and established benchmarks will be needed to evaluate the suitability for different types of problems.

The formal assumptions as presented in this paper may appear to rule out application to many interesting problems such as time series or other situations with non-independent variables. However, since the method is based on locality it seems to be effective in these cases as well, and we are working on a formal generalisation of the statement to cover such cases.

Since performing an exhaustive search over all possible input selections quickly becomes unwieldy, it would be desirable to consider interactions between the Delta test and other search/optimisation schemes. The approach could eventually lead to an effective methodology to complement current methods in use.

References

[1] Roberto Battiti. Using the mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 5(4):537–550, 1994.

[2] M. Verleysen, D. François, G. Simon, and V. Wertz. On the effects of dimensionality on data analysis with neural networks. In *Artificial Neural Nets Problem solving methods*, Lecture Notes in Computer Science 2687, pages II105–II112. Springer-Verlag, 2003.

[3] D. François. *High-dimensional data analysis: optimal metrics and feature selection*. Ph.d. thesis, Université catholique de Louvain, Louvain-la-Neuve, Belgium, 2007.

[4] D. François, V. Wertz, and M. Verleysen. The concentration of fractional distances. *IEEE Transactions on Knowledge and Data Engineering*, 19(7):873–886, 2007.

[5] H. Pi and C. Peterson. Finding the embedding dimension and variable dependencies in time series. *Neural Computation*, 6(3):509–520, 1994.

[6] D. Evans. *Data-derived estimates of noise for unknown smooth models using near neighbour asymptotics*. Ph.d. thesis, Cardiff University, 2002.

[7] E. Liitiäinen and A. Lendasse. Variable scaling for time series prediction: Application to the ESTSP'07 and the NN3 forecasting competitions. In *IJCNN 2007, Orlando, FL, USA*, pages 2812–2816, August 2007.

[8] A. J. Jones. New tools in non-linear modelling and prediction. *Computational Management Science*, 1(2):109–149, 2004.

[9] V. Spokoiny. Variance estimation for high-dimensional regression models. *Journal of Multivariate Analysis*, 82(1):111–133, 2002.

[10] E. Liitiäinen, F. Corona, and A. Lendasse. Nearest neighbor distributions and noise variance estimation. In *ESANN 2007, Bruges (Belgium)*, pages 67–72, April 2007.