A Global Methodology for Variable Selection

Application to Financial Modeling

Qi Yu[*] — Eric Séverin[**] — Amaury Lendasse[*]

[*] *Helsinki University of Technology*
Adaptive Informatics Research Centre
P.O. Box 5400, 02015 Espoo, Finland

qiyu@cc.hut.fi

[**] *Université de Lille I, IAE, Avenue du peuple Belge 104*
59043 Lille, France

ABSTRACT. *In this paper, a global methodology for variable selection is presented. This methodology is optimizing the Nonparametric Noise Estimation (NNE) provided by Delta Test. The 3 steps of the methodology are Forward Selection, Scaling and Projection. The methodology is applies to two examples: the Boston Housing database and a financial data set. It is shown that the proposed methodology provides better input variables than an exhaustive search. Furthermore, interpretability of the results is improved.*

KEYWORDS: *NNE, Delta test, Variable Selection, scaling, projection*

1. Introduction

Variable selection is one of the most important issues in machine learning, especially when the number of observations is relatively small compared to the numbers of variables. It has been the subject in application domains like pattern recognition, process identification, time series modeling and econometrics. In this paper, we focus on its application to the regression problem, in order to discover mathematical relationship between input variables and output variables in the field of finance.

The necessary size of the data set increases exponentially with the number of observations. To circumvent this, one solution is to select the features or variables which best describe the output variables (targets) [Ver 01]. Then, it is possible to capture and reconstruct the underlying regularity or relationship (that is approximated by the regression model) between input variables and output variables.

There are many ways to deal with the feature selection problem, a common one is using the generalization error estimation. In this methodology, the set of features that minimizes the generalization error are selected using Leave-one-out, Bootstrap or other resampling technique [Len 03][Efr 93]. These approaches are very time consuming and may lead to an unacceptable computational time.

However, there are other approaches. In this paper, we use a method called Nonparametric Noise Estimation (NNE), which selects features based only on the dataset. It is then not necessary to build a regression model in order to find the best input variables.

In this paper, NNE is presented in Section 2. Section 3 and 4 describe a global methodology to perform the variable selection using Delta Test. In Section 5, we show some experimental results on a toy example and a financial data set.

2. Nonparametric Noise Estimator using the Delta Test

Delta Test (DT) is a technique for estimating the variance of the noise, or the mean square error (MSE), that can be achieved without overfitting [Jon 04]. The evaluation of the NNE is done using the DT estimation introduced by Stefansson.

Given N input-output pairs : $(x_i, y_i) \in \mathbf{R}^M \times \mathbf{R}$, the relationship between x_i and y_i can be expressed as :

$$y_i = f(x_i) + r_i, \qquad [1]$$

where f is the unknown function and r is the noise. The Delta Test estimates the variance of the noise r.

The DT is useful for evaluating the nonlinear correlation between two random variables, namely, input and output pairs. The DT has been introduced for model selection but also for variable selection : the set of inputs that minimizes the DT is the one that is selected. Indeed, according to the DT, the selected set of variables is the one that represents the relationship between variables and output in the most deterministic way.

DT is based on hypotheses coming from the continuity of the regression function. If two points x and x' are close in the input space, the continuity of regression function implies the outputs $f(x)$ and $f(x')$ will be close enough in the output space. Alternatively, if the corresponding output values are not close in the output space, this is due to the influence of the noise.

Let us denote the first nearest neighbor of the point x_i in the set $\{x_1, \ldots, x_N\}$ by x_{NN}. Then the delta test, δ is defined as :

$$\delta \quad = \quad \frac{1}{2N} \sum_{i=1}^{N} \left| y_{NN(i)} - y_i \right|^2 , \qquad [2]$$

where $y_{NN(i)}$ is the output of $x_{NN(i)}$. For the proof of the convergence of the Delta Test, see [Jon 04].

3. Variable Selection and Delta Test

3.1. *Variable Selection*

The original variable selection problem is to select the k most relevant input variables from a set of d variables $(d \gg k)$. In this paper, the aim of our variable selection is to minimize $Var(r)$ (estimated by Delta test) by selecting k which is unknown. So, what we do is to test all $k = 1, 2, \ldots, k$ and select the one that gives the minimum value of $Var(r)$.

There are several methods for solving both the problem of selecting the optimal number k and the best variables subset. These approaches are introduced in the following sections.

3.2. *Exhaustive search*

The optimal algorithm is to compute the minimum $Var(r)$ for all the possible combinations of input variable. $2^d - 1$ variable combinations are tested (d is the number of input variables). Then, a subset which gives the minimum value of $Var(r)$ is selected.
This procedure is too time consuming and usually, it is impossible to do an exhaustive search. Thus, some faster methods have to be used instead. In the next section, we introduce a global methodology that perform the variable selection in a reasonable time.

4. A Global Methodology

In order to select good input variables in a short computational time, we propose a global methodology in 3 steps.

– Firstly, the most important input variables are selected using a Forward Selection. This initial selection is not optimal but it reduces the number of initial input variables. These initial step can be improved using Backward Selection or Forward-Backward Selection.

– Secondly, a scaling of the variable is performed. This allows the ranking of the input variables. Furthermore, this step improved the performance of symmetric nonlinear models like SVM, LS-SVM or RBFN.

– Thirdly, new variables are build using a linear projection. This step is not mandatory because it reduces the interpretability of the final variables. Nevertheless, it is shown in the experimental section that this last step improves the performances of the global methodology.

The next block diagram summarizes the global methodology. In the 3 steps, the criterion that is optimized is the Delta Test. Forward Selection, Scaling and Projection are presented in the next subsections.

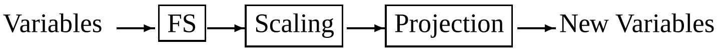

Figure 1. *Block diagram of the global methodology*

4.1. *Forward selection(FS)*

In this method, starting from the empty set S of variable variables, the best variable variable is added to the set S one by one, until the size of S is d (dimension of the variables). Let's suppose that we have a set of variables $\{x_i, y_i\}, i = 1, 2, \ldots, M$, where $x_i \in R^d$, the algorithm is as follows :

1) Set F to be the original set of d variables, and S to be the empty set which will contain the selected variables.

2) Find :

$$x^s = \arg\min_{x^i} Var(r) \quad x^i \in F \qquad [3]$$

where x^s represents the selected variable.
Save the $Var(r)^s$ value and move x^s from F to S.

3) Continue with the same way till the size of S is d .

4) Compare the $Var(r)$ value for all the sizes of sets S, the selected result is set S that minimizes $Var(r)$.

4.2. *Scaling*

Obviously, the input variables can have different importance with respect to the output. Therefore, the data are to be preprocessed and scaled. If variable x is a set of d variables and y is the corresponding output. Then, we can get $Y = f(a_1x^1, a_2x^2, \ldots,$ $a_dx^d)$ using the method of scaling. Variable selection is a particular case of scaling with all the weights $a_i, i = 1, 2, \ldots, d$ equal to 0 or 1. The scaling is important especially if f is a symmetric model like RBFN, SOM or LS-SVM. Here, we search for the best set of weights a_i that minimizes the variance of the noise. I will express that some variables have more importance than others. Small weight in the result indicate that the variable is more irrelevant and weight zero shows the variable can be pruned. Moreover, the variable dimension is determined according to the scaling factors. Of course, this method may be time consuming and requires prior knowledge which is often not available. Thus, in this paper, we use a genetic algorithm to perform the scaling. This method is proved to be efficient as shown in the experimental section.

4.3. *Projection*

In linear algebra, a projection is a linear transformation P, an idempotent transformation. An $d \times k$ matrix projection maps an d-dimensional vector space onto a k-dimensional subspace $(k \ll d)$; such a matrix is also called an idempotent matrix. After projection we get :

$$X_{P(M\times k)} = X_{(M\times d)} \times P_{(d\times k)} \qquad [4]$$

where the original variable space is written as k-dimensional subspace. That means we get k linear combinations of original variable variables as the new variables.

5. Experimental Results

In the experimental section, we are testing the global methodology on 2 databases : one toy example (Boston Housing database) and one real example form the field of finance. It is shown that the global methodology improves the performances but also the interpretability of the results.

5.1. *Boston housing data*

We use a dataset, called Boston housing data [1], which has 506 samples, 13 input variables and one output variable. Then, the methodology presented in Section 4 is used. Obviously, exhausted search for this housing data is very time consuming but it is presented for comparison purposes. Forwards selection method appears to be efficient. The scaling decreases the estimate of Delta Test by 25% and the projection by 30%. The results are illustrated in Table 1. In this example, the dimension projection has been determined by trial and errors. Results with a dimension projection between 5 and 9 are very similar. We have then selected $k = 5$ in order to reduce as much as possible the number of selected variables.

Methods	$Var(r)/Var(y)$	Variables selected
Exhausted search	0.0710	1 3 5 6 7 8 9 10 12 13
Forward selection	0.0755	13 6 1 10 7 11 5 12 2
FS+Scaling	0.0572	
FS+Scaling+Projection	0.0529	Projection Matrix $p_{(9\times5)}$

Tableau 1. *Normalized result comparison (13 variables)*

5.2. *A data set from financial field*

5.2.1. *Results*

In this experiment section, we use a dataset related to 200 French companies during a period of 5 years. 42 input variables are used, these input variables are financial indictors that are measured every year (for example debt, number of employees, amount of dividends, ...). The target variables are

– The ROA defined as the ratio between the net income and the total assets.

– The Marris (or Q ration) is calculated by dividing the market value of shares by the book value of shares

Table 2 shows the real meaning in financial field about all the variables we have used.

Then, we test our global methodology in order to minimize the variance of the noise using Delta test. The variables and results are listed in next figure and Table 3.

The next figure presents the forward selection results including the selected order and the $Var(r)$ value of every selected combinations. Take the figure in left side for example, if we choose 5 on X Label, that means the fifth variables we selected is the

[1] ftp ://ftp.ics.uci.edu/pub/machine-learning-databases/housing/

12nd and the corresponding value on Y Label is the $Var(r)$ value we estimate using the 34th, 26th, 35th, 18th and 12nd variables.

Table 3 shows the normalized $Var(r)$ value we estimated in the 3 steps of the global methodology. From these results, it is quite evident that forward selection is efficient. It is also showing that Scaling and Projection are improving the performances. However, the selection of the projection dimension k has to be performed carefully. Anyhow, we found that setting k around the number of variables selected by the Forward Selection gives satisfactory result. In oder to show the advantages of this global methodology, a financial interpretation of the results is given in the next subsection.

5.2.2. *Interpretation of the results :*

Our sample was chosen from 200 French industrial firms listed on the Paris Bourse (nowadays Euronext) during 1991-1995. We selected all item in balance sheet and income statement able to explain corporate performance. First of all, we try to define precisely corporate performance. An adequate performance indicator should be able to take into account all the consequences on the wealth of stakeholder. We chose ROA and Marris. The first allows measuring the global corporate performance. Nevertheless, this is an "ex-post indicator" because we use book value. That's why we use Marris (or Q ratio). The use of market value allows measuring the future growth opportunities of firm. To sum up, it is an indicator of value creation.

Figure 2 highlights that the first ten variables (for ROA) and the first fifteen variables (for Marris) are the best combination to explain performance. When we exceed this edge, we deteriorate the performance. The new variables do not allow improving the explanation of corporate performance.

For ROA and Marris, Figure 2 shows that size variable have a positive influence on performance. Indeed, the size can give a power market able to have best performance. For instance we have the possibility to put pressure on customers and suppliers. More interestingly, figure 2 highlights that leverage has a positive influence on Marris (i.e. on the future growth opportunities). This result is in line with "free cash flow" hypothesis. In this framework debt is the best way to reduce conflicts of interest. Leverage provides discipline and monitoring not available to a firm completely financed by equity. According to the "free cash flow" theory, debt creates value by imposing discipline on organizations which in turn reduces agency costs. The use of debt has two functions : 1) it decreases the free cash flow that can be wasted by managers and, 2) it increases the probability of bankruptcy and the possibility of job loss for managers (thus leading to the disciplining effect).

To continue the further analysis of the these variables, we use the result of $FS + Scaling$ method. Table 4 shows the variables selected by FS method and the scaling factors on them for Output 1 and Output 2.

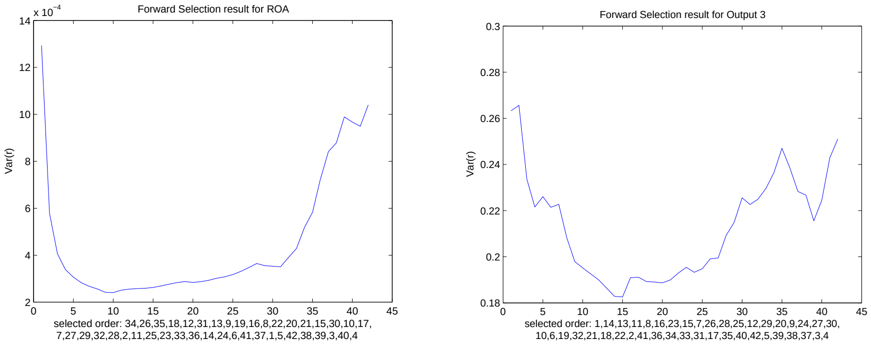

Figure 2. *Forward selection figure for financial data*

6. Conclusion

In this work, we use non-parameter technique to perform variable selection. Despite the fact that the model f is unknown, the Delta test privides an estimate for the variance of the noise $Var(r)$.

According to the experimental results, it is shown that the global methodology that we have proposed gives better performance than an exhaustive search. Furthermore, it is then possible to give interpretation to the selected variables.

In further work, better method for the selection of the projection dimension k will be be investigated and the global methodology will be tested on new data especially from the field of finance.

7. Bibliographie

[Ver 01] VERLEYSEN M., « Learning high-dimensional data », *NATO Advanced Research Workshop on Limitations and Future Trends in Neural Computing*, Italy, 2001, p. 22-24.

[Len 03] LENDASSE A.,WERTZ V., VERLEYSEN, M., « Model Selection with Cross-Validations and Bootstraps - Application to Time Series Prediction with RBFN Models », *Joint International Conference on Artificial Neural Networks*, Istanbul, Turkey. 2003

[Efr 93] EFRON B., TIBSHIRANI R., « An Introduction to the Bootstrap », *Chapman and Hall*, London, 1993

[Jon 04] JONES A. J., « New Tools in Non-linear Modeling and Prediction », Computational Management Science, n^{o} 1, 2004, p. 109–149.

index	Variable	Meaning
1	SECTEUR	Industry
2	Transaction	Number of shares exchanged during the year
3	Rotation	Security turnover rate
4	VÃ©rif Rotation	Not useful
5	Dividende net	Amount of dividend for one share during the year
6	Effectifs	Number of employees
7	CA	Sales
8	II	Other assets
9	AMORII	Dotations on other assets
10	IC	Property, plant and equipement
11	AMORIC	Dotations on property, plant and equipement
12	IF	Not useful
13	AI	Fixed assets
14	S	Stocks or inventories
15	CCR	Accounts receivables
16	CD	Not useful
17	L	Cash in hands and at banks
18	AC	Total of current assets
19	CPPG	Total of capital of group (in book value)[a]
20	PRC	Not useful
21	FR	Accounts payables
22	DD	Not useful
23	DEFI	Financial debt
24	DETTES-1AN	Debt whose maturity is inferior to 1 year
25	DETTES+1AN	Debt whose maturity is superior to 1 year
26	TD	Total Debt
27	CA	Sales
28	CPER	Cost of workers
29	CPO	Not useful
30	DA	Dotations on amortizations
31	REXPLOI	Operating income before tax
32	CFI	Interests taxes
33	RFI	Financial income
34	RCAI	Operating income before tax + Financial income
35	REXCEP	Extraordinary item
36	IS	Taxes from State
37	RPI	(II+IC+AMORIC+IF)/TA :the renewal policy of the immobilizations after investments
38	AI	AMORIC/(II+IC+AMORIC+IF) :measures the age of the immobilizations
39	FA	TD/TP :measures the financial autonomy
40	DM	(DETTES+1AN)/TD : debt maturity
41	LC	CPER/CA : labor cost in the company
42	FDR	(S+CCR-FR)/CA : working capital of the company
43	ROA	net income / total assets
44	MARRIS	Market to book

[1] By construction the total debt is equal to Total assets

Tableau 2. *The meaning of variables*

	$Output_1$ (ROA)	$Output_2$ $(MARRIS92)$
Forward and variables selected	0.1001 (34,26,35,18,12 31,13,9,19,16)	0.4965 (1,14,13,11,8,16,23,15 7,26,28,25,12,29,20)
Scaling with the variables selected by Forward	0.0746 first 10 variables	0.4690 first 15 variables
Forward+Scaling +Projection	0.0635 $P_{(10\times 10)}$	0.3475 $P_{(15\times 5)}$

Tableau 3. *Normalized Delta test result with variables selected*

(a) Output 1 : ROA

index	Variable	Scaling value
26	TD	0.9996
34	RCAI	0.9976
31	REXPLOI	0.9666
19	CPPG	0.8167
13	AI	0.7711
9	AMORII	0.7552
35	REXCEP	0.5367
12	IF	0.2872
18	AC	0.2685
16	CD	0.1468

(b) Output 2 : Marris

index	Variable	Scaling value
23	DEFI	0.9975
28	CPER	0.9689
20	PRC	0.9003
11	AMORIC	0.8940
12	IF	0.8849
25	DETTES+1AN	0.8501
26	TD	0.8252
15	CCR	0.6995
7	CA	0.6968
8	II	0.5966
14	S	0.5705
29	CPO	0.4368
16	CD	0.3732
13	AI	0.3529
1	SECTEUR	0.1049

Tableau 4. *The Scaling factors for the selected variables*