Variable Selection for Financial Modeling

Qi Yu[1], Eric Séverin[2] and Amaury Lendasse[1]

[1] Helsinki University of Technology, Adaptive Informatics Research Centre,
P.O. Box 5400, 02015 Espoo, Finland
[2] 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 Variable Selection (VS), 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.

1 Introduction

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

There are many ways to deal with the feature selection problem. In this paper, a new method called Nonparametric Noise Estimation (NNE) is used. The NNE method selects the best features and simultaneously estimates a priori the performance of the best nonlinear regression model that can be built with the selected variables. A methodology based on Nonparametric Noise Estimation is presented. The successive steps of the methodology are: 1. The selection of the most important input variables. 2. To compute the optimal weight of each selected variable in the regression problem. Each weight is in a range between 0 and 1. 3. New variables that improve performances are built. These new variable are linear combinations of the previously selected ones.

In the experimental section, from a sample of 200 French companies during a period of 5 years, we use 42 variables (inputs) coming from balance sheet, income statement and market data. We seek to highlight the main variables able to explain some financial indicators (the outputs or targets variables): ROA, ROE and Marris (or Tobins Q). The first two indicators have been retained because they are economic and financial indicators of corporate efficiency. The last is a means to measure the value creation for shareholders because it represents the future

growth opportunities of the firms. Thus, by using this methodology, we can select, ranked and scaled the inputs. For the example of ROA, 9 input variables can be retained. It is also possible to build 5 new variables that improve the regression performances. For the ROA, the estimation of the final performances gives a correlation coefficient (R2) equal to 0.94. To sum up, the presented methodology allows the determination of the most important variables for nonlinear regression problems. The main advantage of the methodology is that the nonlinear model does not need to be computed. Furthermore, it is shown that the methodology is efficient even if the number of data (samples) is very small. In further work, new examples of financial regression problems will be studied.

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 [4]. 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:

$$y_i = f(x_i) + r_i, \tag{1}$$

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:

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

where $y_{NN(i)}$ is the output of $x_{NN(i)}$. For the proof of the convergence of the Delta Test, see [4].

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 $k = 1, 2, \ldots, d$ 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 set of input variables are selected using a Forward-Backward Selection (FBS). This initial selection is not optimal but it reduces the number of initial input variables.
- Secondly, a scaling of the selected variables is performed. This shows the different weights of these variables and 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-Backward Selection (FBS), Scaling and Projection are presented in the next subsections.

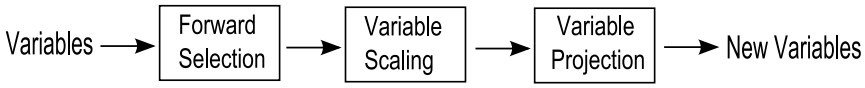

Fig. 1. Block diagram of the global methodology

4.1 Forward-Backward selection(FBS)

Forward-Backward Selection combines both Forward and Backward Selection, and it avoids the incomplete search suffered from using them seperately. This method can be started from any of the input, even randonly initialized set. Let's suppose that we have a set of variables $\{x_i, y\}, i = 1, 2, \ldots, M$, where $x_i \in R^d$, the algorithm is as follows:

1. Set F to be the initial input set which can contain any input variables, and set S to be the unselected input set containing the rest of input variables which are not in set F. Compute the $Var(x^F, y)$ using input variables in set F and the output y
2. Find:

$$x^F = \arg\min_{x_{i,j}} \{Var(x^F \cup x_i, y), Var(x^F \setminus x_j, y)\} \quad x^i \in S, x^j \in F \qquad (3)$$

   where x^F represents the selected variables set.

3. If the old value of $Var(x^F, y)$ is smaller than the new result, stop; otherwise, update set F and save the new result, repeat the step 2 till no further change can decrease the Variance result.
4. The final selected result is in set F.

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. The scaling is important especially if the function is a symmetric model like RBFN, SOM or LS-SVM. Variable scaling can be seen as a generalization of variable selection; instead of restricting the scalars to attain either values 0 or 1, the entire range $[0, 1]$ is allowed.

In this section, we present a method for choosing the scaling using Delta Test (DT). FBS is also extended to scaling as well. Instead of turning scalars from 0 to 1 or vice versa, in this method,increasing by $1/h$ and decreasing by $1/h$ are allowed. Integer h is a constant grid parameter. Thus, 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.

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:

$$X_{P(M\times k)} = X_{(M\times d)} \times P_{(d\times k)} \tag{4}$$

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

Thus, the new variables are built from the k linear combinations of original variables. Especially, we investigate a method which can automatically find the best projection dimension k to minimize the Variance value. The results are shown in the following section.

5 Experimental Results

In the experimental section, we are testing the global methodology on 2 databases: one example is Boston Housing database and the other one is 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 [5], which has 506 samples, 13 input variables and one output variable. Then, the methodology presented in Section 4 is used. Obviously, FBS method appears to be efficient. The VS+Scaling+Projection decreases the estimate of Delta Test by 20%. The results are illustrated in Table 1.

Table 1. Normalized result comparison (13 variables)

Methods	$Var(r)/Var(y)$	Variables selected (weight)
Variable Selection	0.1055	3 6 9 13
VS+Scaling	0.0890	3(0.4) 6(0.8) 9(0.4) 13(0.4)
VS+Scaling+Projection	0.0864	Projection Matrix $p_{(4\times 4)}$

5.2 A data set from financial field

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 ROE represents the ratio between the net income and the capital
- 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 results are listed in the following Table 3.

To continue the further analysis of the these variables, we use the result of $FS + Scaling$ method. Table 4, Table 5and Table 6 shows the variables selected by FS method and the scaling factors on them for Output 1, Output 2 and Output 3.

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 items 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, ROE 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.

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).

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 prividesan 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.

Table 2. The meaning of variables

index	Variable	Meaning
1	Sector	Industry
2	Transaction	Number of shares exchanged during the year
3	Rotation	Security turnover rate
4	Vrif Rotation	Not useful
5	Net dividend	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	Debt-1AN	Debt whose maturity is inferior to 1 year
25	Debt+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	ROE	net income / capital
45	MARRIS	Market to book

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

Table 3. Normalized result

	$Output_1$ (ROA)	$Output_2$ (ROE)	$Output_3$ ($MARRIS92$)
Variable Selection (VS) (selected amount)	0.0999 (9)	0.2551 (10)	0.5377 (19)
VS+Scaling	0.0751	0.1879	0.4683
VS+Scaling+Projection Projection Matrix P	0.0595 $P_{(9\times5)}$	0.1585 $P_{(10\times7)}$	0.4115 $P_{(19\times11)}$

Table 4. Output 1: ROA

index	Variable	Scaling value
34	RCAI	1.0
26	TD	0.8
19	CPPG	0.7
31	REXPLOI	0.7
9	AMORII	0.6
35	REXCEP	0.4
16	CD	0.4
12	IF	0.3
13	AI	0.3

Table 5. Output 2: ROE

index	Variable	Scaling value
31	REXPLOI	1.0
34	RCAI	1.0
13	AI	1.0
37	RPI	0.8
36	IS	0.7
24	Debt-1AN	0.7
35	REXCEP	0.6
33	RFI	0.3
1	Sector	0.3
19	CPPG	0.2
39	FA	0.1
32	CFI	0.1
30	DA	0.1
18	AC	0.1
26	TD	0.0
14	S	0.0
9	AMORII	0.0
5	Dividende net	0.0

References

1. Verleysen, M.: Learning high-dimensional data. NATO Advanced Research Workshop on Limitations and Future Trends in Neural Computing. Italy, (2001) 22–24

Table 6. Output 3: Marris

index	Variable	Scaling value
28	CPER	1.0
22	DD	1.0
29	CPO	0.9
19	CPPG	0.9
15	CCR	0.8
11	AMORIC	0.8
7	CA	0.8
41	LC	0.7
13	AI	0.7
25	Debt+1AN	0.5
5	Net dividend	0.5
42	FDR	0.2
39	FA	0.2
24	Debt-1AN	0.2
9	AMORII	0.2
40	DM	0.1
34	RCAI	0.1
2	Transaction	0.1
1	Sector	0.1

2. 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)
3. Efron, B., Tibshirani, R.: An Introduction to the Bootstrap. Chapman and Hall. London, (1993)
4. Jones, A. J.: New Tools in Non-linear Modeling and Prediction, Computational Management Science, (1), (2004), 109–149
5. ftp://ftp.ics.uci.edu/pub/machine-learning-databases/housing/