Multistart Strategy Using Delta Test for Variable Selection

Dušan Sovilj

Aalto University School of Science
Konemiehentie 2, Espoo, Finland
dusans@cis.hut.fi

Abstract. Proper selection of variables is necessary when dealing with large number of input dimensions in regression problems. In the paper, we investigate the behaviour of landscape that is formed when using Delta test as the optimization criterion. We show that simple and greedy Forward-backward selection procedure with multiple restarts gives optimal results for data sets with large number of samples. An improvement to multistart Forward-backward selection is presented that uses information from previous iterations in the form of long-term memory.

Keywords: Delta test, noise variance, long-term memory, restart strategy, variable selection, forward-backward selection.

1 Introduction

The number of features or attributes in newly available data sets grows rapidly mostly due to easier data acquisition, storage and retrieval. When the problem is regression, i.e. predicting the real value for a fresh sample, most machine learning methods use all available features which can degrades the predictions [1]. Therefore, proper selection of variables is needed before training the model. Benefits of variable selection are twofold: increasing prediction accuracy and interpretability.

One of the criteria used for variable selection is the Delta test, a noise variance estimator as proposed in [2]. In order to select an optimal set of variables, one should examine an exponential number of possibilities which depends on the dimensionality of the data set. In the case of Delta test, certain variables subsets may be ignored due to the nature of the criterion as explained in later sections.

Forward-backward selection [3] is a simple and widely used procedure for variable selection. The idea is to either include or exclude a single variable at a time and compute the criterion for the obtained subset of variables and repeat the process until the criterion does not improve. This procedure is sufficient when using Delta test and when there are enough samples, but it requires several restarts from random subsets to reach a satisfying solution. When restarting is involved a lot of information about the search process can be used to influence the later stages. This information is reused in some search algorithms, such as

T. Honkela et al. (Eds.): ICANN 2011, Part II, LNCS 6792, pp. 413–420, 2011.

Tabu search [4] and variants of Greedy randomized adaptive search procedures (GRASP) [5][6].

The paper is organized as follows. Section 2 explains the Delta test criterion. In Section 3 Forward-backward selection is briefly mentioned. Section 4 explains the idea of multistart strategies and a specific implementation for Delta test optimization. The results of experiments are given in Section 5 and finally in Section 6 the concluding remarks are presented.

2 Delta Test

Delta test (DT) is a non-parametric noise variance estimator based on a nearest neighbour principle. The estimator is used when a functional dependence is assumed between inputs $\mathbf{x}_i$ and output y_i with additive noise term, i.e. $y_i = g(\mathbf{x}_i) + \epsilon_i$, given finite number of samples $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, $i = 1, \ldots, M$. The function g is assumed to be smooth, and the noise terms ϵ_i are i.i.d. with zero mean. The Delta test estimates the variance with the following formula:

$$\mathrm{Var}(\epsilon) \approx \frac{1}{2M} \sum_{i=1}^{M} (y_i - y_{NN(i)})^2 \,, \tag{1}$$

where $NN(i)$ defines the nearest neighbour of sample $\mathbf{x}_i$ in the input space. In [7] it is shown that the estimator converges to the true value of the noise variance in the limit $M \to \infty$.

As shown in [2], DT requires many samples in order to find correct subset of variables in a noisy setting. On the other hand, when presented with an adequate sample pool, it is able to distinguish between important and noisy variables in most subsets (Section 5.1). In this case, Forward-backward selection is a potent procedure to look for the optimal selection.

The case with an inadequate number of samples is an ill-posed situation. Although DT overestimates the noise variance in this case [7], the selection with the optimal DT selection may also include noisy variables. Therefore, the search should focus on subsets that contain small number of variables to more precisely estimate the variance, while the subsets with more variables can be completely ignored.

The aim in variable selection is to find a subset of variables that minimizes DT estimate. Previous work in this domain [3] is oriented on comparing different search algorithms on data sets with small number of samples and with a time constraint. The main focus in this paper is shape of Delta test landscape when we have enough samples.

3 Forward-Backward Selection

Forward-backward selection (FBS) is a simple procedure for variable selection involving any criterion in machine learning. For DT optimization, the procedure starts from any solution s, i.e. any non-empty subset of variables. Then

it evaluates all *neighbours* $N(s)$ (solutions with either 1 more or 1 less variable than in s) and picks the subset with the smallest DT value from $N(s)$ as the *new* s. These two phases (evaluation and selection) are repeated until no further improvement is possible.

4 Multistart/Restart Strategies

Restarting phase to improve results has been of interest in different search algorithm domains [5][10][11]. Our focus is on simple construction of a solution, from which FBS converges to a local minima. Such approach uses *long-term memory*, a structure that originates from Tabu search [4]. The idea in these search procedures is to construct a high quality solution from which greedy algorithm requires less steps than from a random starting position.

Long-term memory structures keep track of certain aspects of solutions encountered during various stages of the search. Solutions that are good in terms of criterion (also called objective function) are called elite solutions, and most of long-term information relates to elite solutions. Information gathered usually involves: frequency of a variable residing within elite solutions; and the impact of a variable change on objective function value. The former is called consistency and the latter strong determination [8]. This information is used in the first phase of the approach – the construction phase used to build the starting position. Each piece of information contributes to the *energy* of a variable $E(i)$, $i = 1, \ldots, d$ with d being the dimensionality of the data set. Construction is done by adding single variable at a time, usually in a probabilistic setting. The probabilities are obtained from energies in the usual way, i.e. $p(i) = E(i)/\sum_{i=1}^{d} E(i)$. After the construction phase, greedy algorithm is used to find the local optimum. Construction phase and convergence constitute one iteration of the approach.

4.1 Multistart Strategy for Delta Test

For the experiments, both consistency and strong determination are used to guide the construction phase towards promising starting solutions. After obtaining the starting position, the FBS is used for descent phase.

For the rest of the paper we use the following notation: s is the solution or subset of variables; $f(s)$ is DT estimate using the solution s; $f_{\min}$ the smallest estimate found during search; $C(i)$ is consistency of a variable i; $S(i)$ is strong determination of variable i; S_v is the number of variables to add in construction phase; s_e an elite solution; S_e a set of elite solutions (elite memory); and s^{i-} and s^{i+} indicate that i-th variable is excluded and included in solution s respectively.

Energy function is defined as

$$E(i) = \lambda S(i) + C(i)\,, \quad i = 1, \ldots, d\,, \tag{2}$$

where λ controls the trade-off between the two terms and for the experiments we set λ to give equal weight to both terms.

Consistency for a variable i is computed as

$$C(i) = \sum_{s_e^{i+}} \frac{f_{min}}{f(s_e)}\,, \tag{3}$$

i.e. the sum of DT ratios between best solutions found during the search and elite solutions that have i-th variable included. Consistency tells how frequent is a variable in elite solutions with higher values indicating that a variable is more important.

Strong determination is computed as

$$S(i) = \frac{1}{K} \sum_{j=1}^{K} \frac{f(s_j^{i-})}{f(s_j^{i+})}\,, \tag{4}$$

where the fraction denotes the change in objective function for each solution s_j when i-th variables is included, and K the number of such fractions. The higher the value of the fraction in the sum indicates that variable i should be included in the current solution keeping all other variables intact. During the descent phase with FBS, each variable is flipped to check the neighbouring solutions for improvement. These flips also enable us to compute the strong determination for all variables since both $f(s_j^{i-})$ and $f(s_j^{i+})$ are available.

We take only the last 3 changes of descent phases where new local minima are discovered, since those are more important ones that contribute to the f around local minimum. Also, the first iterations have longer descents than later ones as better solutions are generated in later iterations, and we want to treat all iterations equally. The idea of strong determination is too find variables which are good no matter what area of optimization landscape the search algorithm is exploring.

Since $C(i)$ depends on the number of elite solutions, it is at most $|S_e|$. To make both terms equal in Equation (2), λ is set to $|S_e|$ and all values of $S(i)$ are divided by the maximum among $S(i)$. Thus, both terms have roughly the same magnitude when computing energy.

Number of variables. Construction phase selects S_v variables based on the following procedure. For S_v passes a variable is selected probabilistically, but only certain number of best variables (not included in the partial solution) are kept for the selection step. Thus, we first take $P\%$ of the largest $E(i)$ and then select one variable based on $p(i)$. The energy function usually includes the value of objective function $f(s^{i+})$ as an extra term, but in the case of Delta test it is not needed as consistency and strong determination are sufficient to guide the construction toward good solutions making the building phase much quicker.

In order to properly explore the solution space, a diversification strategy is needed. One of the ways to achieve this is by changing the P value. With smaller values, the focus is only on good variables and the generated solutions to not differ too much. On the other hand, with $P \approx 1$ the generated solutions are too

diverse and gathered information is not properly exploited. In the experiments, we set P to a constant 0.5 value during complete search process. It is a good value for tested data sets without hindering the exploration. More refined strategies change P based on diversity of generated solutions with some kind of measure [9]. As mentioned, a constant value of 0.5 makes a good compromise that does not involve any complex strategies with additional parameters.

The correct number of variables before the search is unknown, but since DT estimates noise variance more precisely with less variables, the desired goal is to obtain solution s with smallest number of variables and still keep estimate $f(s)$ minimized. Therefore, parameter S_v should be initiated to a small value and adjusted as the search progresses. In each iteration, S_v is made equal to the number of variables in the best solution found during search. The idea is to have at most variables as in the best solution, and still focus on minimizing f. Of course, more complex approaches are possible.

Diversity of solutions. The consistency value depends on elite solutions s_e in memory. To be able to produce diverse starting solutions, the elite ones have to be diverse enough themselves. Therefore, we define solution s an elite solution if:

1. $f(s) < f_{\min}$ i.e. it has the smallest DT estimate
2. $f(s) < f(s_{e_k})$ for some s_{e_k} and s is sufficiently diverse from better solutions $\{s_{e_j} | f(s_{e_j}) < f(s), j = 1, \ldots, k-1\}$

The diversity is taken as the percent of different values of variables, but only on those positions where the variables are set to 1. If $D_{s,t} = \{i \,|\, s^{i+} \vee t^{i+},\ i = 1, \ldots, d\}$ defines a set of variables which have value 1 in s or t, then solutions s and t are diverse if $\sum_{i \in D_{s,t}} s^i \neq t^i \ \geq \ \max(2, |D_{s,t}|/4)$, i.e. solutions s and t have to *disagree* on at least quarter of selected variables combined together. If the solution s enters the elite memory, then all similar solutions t with $f(t) > f(s)$ are removed from elite memory ensuring diversity over entire memory range.

5 Experiments

5.1 Synthetic Data

In order to see how good FBS is, consider a simple artificial data set as a function of three variables $f(x_1, x_2, x_3) = \cos(2\pi x_1)\cos(4\pi x_2)\exp(x_2)\exp(2x_3) + \epsilon$ with signal-to-noise ratio close to 1 (as used in [2]) and with increasing number of completely irrelevant ones. Figure 1 shows the influence of the number of samples and dimensions on the number of local minima and the optimization landscape. More samples makes the valley around global minima more steeper and "wider" enabling FBS to reach optimal solution from most starting points. When the number of samples is low, the global minima does not necessary correspond to correct selection of variables. Nevertheless, even in such a scenario, using FBS provides that global solution from certain number of starting positions which

decreases with increasing number of dimensions. The number of samples clearly separates the problem into two categories, and thus different algorithms are needed depending on the situation.

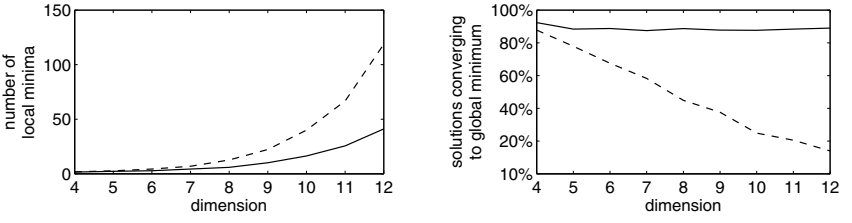

Fig. 1. Average number of local minima (left) and average percentage of all solutions from which FBS converges to global minimum (right). Results are averages over 100 generated data sets with M=256 samples (dashed line) and M=8192 (solid line).

With high number of samples DT more easily identifies noisy variables from important ones in almost complete optimization landscape. To make sure the global minimum is reached with FBS, several starting positions are needed, but not too many since from most solutions FBS converges to the global one. The next subsection shows results when evaluation of all solutions in not feasible due to the high number of variables.

5.2 Real-World Data

We denote with M_{rnd} the restart strategy with FBS from a random point, and with M_{con} the proposed strategy with construction phase. The strategies are tested on two data sets. One is modified Anthrokids[1] data set with removed missing values, containing 1019 samples in 53 dimensions. The other data set is formed from well known Santa Fe Competition Data – Series A[2], but with training and test parts combined. The data set is formed by having a regressor of size 36 producing 10057 samples.

Figure 2 shows the convergence of both strategies for two data sets. The results are averages over 10 runs for both strategies, and the parameters for Santa Fe are ($S_v = 5, |S_e| = 10$) and for Anthrokids ($S_v = 10, |S_e| = 20$). Construction phase in the first couple of iterations should not heavily favor any variables until enough solutions have been evaluated. Thus, strong determination is slightly altered by adding constant value of 10 to the sum in Equation (4) for the first 5 iterations, after which it is dropped.

From the figure we see that the final value of DT estimate is almost the same given more iterations, but the proposed M_{con} strategy has the advantage of converging to those values faster. For Santa Fe 9 variables are selected giving DT estimate of 7.0107, while for Anthrokids we have 15 variables and estimate

[1] `http://research.ics.tkk.fi/eiml/datasets/Anthrokids.zip`
[2] `http://www-psych.stanford.edu/~andreas/Time-Series/SantaFe.html`

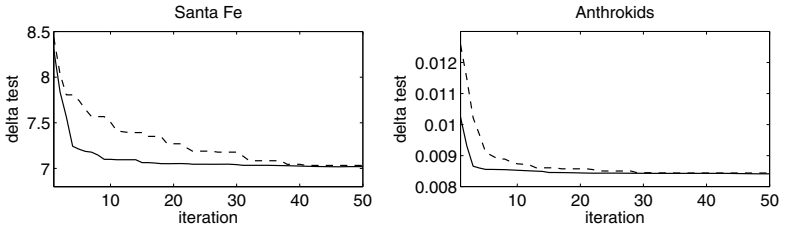

Fig. 2. Convergence of DT as a function of iterations for Santa Fe (left) and Anthrokids (right). Dashed line represents M_{rnd} strategy and solid line M_{con} strategy.

of 0.0085 (normalized output). Given more iterations for Anthrokids (greater than 50), even better solution is possible: 0.0083 with 12 variables.

Figure 3 shows the number of steps per iteration for the same set of experiments. Proposed M_{con} strategy clearly enables better starting position by including long-term memory information. Therefore, a lot less changes are required to converge to a local minima and the total number of DT evaluations is decreased.

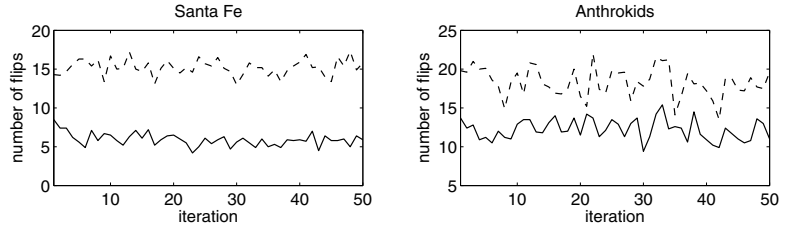

Fig. 3. Average number of steps for FBS per iteration for Santa Fe (left) and Anthrokids (right). Dashed line is M_{rnd} and solid line M_{con} strategy.

One downside of the approach is reliance on FBS which requires examining *all* d neighbours at each step. In situations where d is large, making a single flip can be computationally demanding, as well as one iteration. In this situation, one approach is to fix variables with higher energies before construction phase, or those that are included in all elite solutions.

6 Conclusions

We proposed a multistart strategy for Delta test optimization for variable selection. Due to the nature of landscape that is formed by DT and given the adequate number of samples, the estimator is able to distinguish between noisy and useful variables in most of the optimization landscape. The simple Forward-backward selection procedure is able to reach global minimum given couple of

starting points. To speed up convergence, long-term memory information in form of consistency and strong determination is used. In the experiments, this information showed to be good both in terms of faster convergence and generating solution from which couple of changes are needed to reach local minima.

For further work, a lot more data sets must be tested with large number of samples and without information about relevancy of variables. Proposed strategy includes a lot of parameters, but their meanings should be more intuitive compared to other optimization algorithms and constitute a trade-off between speed of convergence and quality of solutions. For much higher dimensional data sets with over 100 variables, the strategy should be able to provide good results in few iterations.

References

1. Verleysen, M., François, D.: The curse of dimensionality in data mining and time series prediction. In: Cabestany, J., Prieto, A.G., Sandoval, F. (eds.) IWANN 2005. LNCS, vol. 3512, pp. 758–770. Springer, Heidelberg (2005)
2. Eirola, E., Liitiäinen, E., Lendasse, A., Corona, F., Verleysen, M.: Using the Delta Test for Variable Selection. In: European Symposium on Artificial Neural Networks 2008, pp. 25–30 (2008)
3. Guillén, A., Sovilj, D., Mateo, F., Rojas, I., Lendasse, A.: Minimizing the Delta Test for Variable Selection in Regression Problems. International Journal of High Performance Systems Architecture 1, 269–281 (2008)
4. Glover, F., Laguna, F.: Tabu Search. Kluwer Academic Publishers, Norwell (1997)
5. Fernandes, E.R., Ribeiro, C.C.: Using an adaptive memory strategy to improve a multistart heuristic for sequencing by hybridization. In: Nikoletseas, S.E. (ed.) WEA 2005. LNCS, vol. 3503, pp. 4–15. Springer, Heidelberg (2005)
6. Fleurent, C., Glover, F.: Improved Constructive Multistart Strategies for the Quadratic Assignment Problem Using Adaptive Memory. INFORMS J. on Computing 11, 195–197 (1999)
7. Liitiäinen, E., Corona, F., Lendasse, A.: Nearest neighbor distributions and noise variance estimation. In: European Symposium on Artificial Neural Networks 2007, pp. 67–72 (2007)
8. Resende, M.G.C.: Greedy randomized adaptive search procedures (grasp). Technical report, AT&T Labs Research (1998)
9. Morrison, R.W., De Jong, K.A.: Measurement of population diversity. In: Collet, P., Fonlupt, C., Hao, J.-K., Lutton, E., Schoenauer, M. (eds.) EA 2001. LNCS, vol. 2310, pp. 31–41. Springer, Heidelberg (2002)
10. Jansen, T.: On the analysis of dynamic restart strategies for evolutionary algorithms. In: Guervós, J.J.M., Adamidis, P.A., Beyer, H.-G., Fernández-Villacañas, J.-L., Schwefel, H.-P. (eds.) PPSN 2002. LNCS, vol. 2439, pp. 33–43. Springer, Heidelberg (2002)
11. James, T., Rego, C., Glover, F.: Multistart Tabu Search and Diversification Strategies for the Quadratic Assignment Problem. IEEE Transactions on Systems, Man and Cybernetics 39, 579–596 (2009)