Journal of
Ship&Ocean
Technology
©DECEMBER 2007

SOTECH Vol. 11, No. 4, pp. 21 ~ 28, 2007

Enhanced Genetic Programming Approach for a Ship Design

Kyung Ho Lee[1], Young Soo Han[2] and Jae Joon Lee[2]

[1] Department of Naval Architecture and Ocean Engineering, Inha University, Incheon, Korea
[2] Department of Naval Architecture, Graduate School of Inha University, Incheon, Korea; Corresponding Author: kyungho@inha.ac.kr

Abstract

Recently the importance of the utilization of engineering data is gradually increasing. Engineering data contains the experiences and know-how of experts. Data mining technique is useful to extract knowledge or information from the accumulated existing data. This paper deals with generating optimal polynomials using genetic programming (GP) as the module of Data Mining system. Low order Taylor series are used to approximate the polynomial easily as a nonlinear function to fit the accumulated data. The overfitting problem is unavoidable because in real applications, the size of learning samples is minimal. This problem can be handled with the extended data set and function node stabilization method. The Data Mining system for the ship design based on polynomial genetic programming is presented.

Keywords: genetic programming(GP), data mining, ship design

1 Introduction

Although Korean shipyards have accumulated a great amount of data, they do not have appropriate tools to utilize the data in practical works. Engineering data contains the experiences and know-how of experts. Data mining technique is useful to extract knowledge or information from the accumulated existing data. This paper presents a machine learning method based on genetic programming (GP), which can be one of the components for the realization of data mining. The paper deals with polynomial GP for regression or approximation problems when the given learning samples are not sufficient.

Polynomials are widely used in many engineering applications such as response surface modeling(Simpson et al. 1998, Malik et al. 1986, Alotto et al. 1997, Ishikawa et al. 1997, Myers et al. 1995), since the mathematical form of a polynomial is simple, and very easy to handle. In the mechanical, electrical and electronic engineering field, the response surface method is adopted to reduce the computational cost required for analysis and simulation during the optimization design process. Thus, it is desirable to use the minimal size of samples to construct response surfaces. The classical method for attaining good

polynomials is to use "all-possible-regressions" and "stepwise regression" methods(Ott 1993, Myers et al. 1995), but there are limitations in obtaining polynomials with a desired accuracy.

The purpose of this paper is to develop the data utilization tool with genetic programming approach. That is, we focus on the development of Data Mining system to generate an empirical formula from the accumulated existing data for ship design.

In this paper, we try to use genetic programming (GP)(Koza 1992) for generating optimal polynomials that approximate very highly nonlinear response surfaces using only minimal or very small size of learning samples. Major issues regarding finding such polynomials using GP are addressed below.

First, the GP tree can easily represent the polynomial if a function set contains only "+", "-", "*" operators, and a terminal set includes only variables and constants, but it is difficult to expect for genetic programming(GP) to generate polynomials enabling to model a nonlinear function using only such a function and terminal set. We tackle this problem by the use of low order Taylor series of various mathematical functions in a function set. That effectively makes GP produce very high order polynomials. But the generated polynomial tends to become too complex, and it is necessary to control the size of polynomials.

Second, the overfitting problem can be very serious, because we only have small learning samples, and there are no other kinds of additional samples. So, we introduced the EDS(Extended Data Set) method(Yeun et al. 1999) with the FNS(Function Node Stabilization) method. The paper deals with above-mentioned issues, and also presents the Data Miner as a data utilization tool.

2 Genetic programming(GP) for data mining

2.1 Function set for GP

Our idea for easily generating a high order polynomial is to use Taylor series of mathematic functions in the function set. If high order series are taken, the GP tree produces a very complex polynomial. So, we determine to take only two or third order polynomials from whole Taylor series. We use the following function and terminal set.

$$F = \{+,-,*, g_i (i = 1,...,18)\}$$

$$T = \{one, rand, x_i (i = 1,..., n)\}$$

g_i is a low order Taylor series. Its description will not be given in this paper due to the space limitation. "one" return 1, and "rand" a random number, whose size is less than 1. x_i represents a variable. All functions and terminals have their weights. The weights are estimated using the Hooke & Jeeves method towards further minimizing its fitness function defined in the next section. Since we adopt Taylor series, the ordinary least squared method(Myers 1995) cannot be used easily.

2.2 Overfitting avoidance

Without considering overfitting, the fitness function can be defined by (1).

$$\vartheta = \vartheta_{MSE} \tag{1}$$

where $\vartheta_{MSE} = 1/m \sum_{i=1}^{m} [f_{GP}(\overline{X}_i) - y_i]^2$.

The learning set takes the form of $L\{(\overline{X}_i, y_i), 0 \le y_i \le 1\}_{i=1,...,n}$.
Here, $\overline{X}_i(x_{i,1},..., x_{i,n}, 0 \le x_{i,j} \le 1)$ is a n-dimensional vector, and y_i is a desired output of the GP tree(f_{GP}) at $\overline{X}_i$. Since m is very small, the GP tree is overfitted if only $\vartheta = \vartheta_{MSE}$ is used. The EDS method[8] can be included in the fitness function for smooth fitting.

$$\vartheta_E = \vartheta_{MSE} + \lambda \hat{\vartheta}_{MSE} \tag{2}$$

where $\hat{\vartheta}_{MSE} = 1/p \sum_{i=1}^{p} [f_{GP}(\overline{X}_i^E) - y_i^E]^2$, and λ is a constant that determines the contribution of $\hat{\vartheta}_{MSE}$.

The extended data set $L^E\{(\overline{X}_i^E, y_i^E)\}_{i=1,...,p}$ can be constructed by simple linear interpolations of closest learning samples. For the detailed description, see the literature(Yeun et al. 1999). In this paper, we simply take 0.1 as the value of λ. That is a small value, and certainly the GP tree is overfitted. To alleviate this situation, we introduced the FNS method. If the GP tree contains several function nodes $(g_i\ T)$, where T is the subtree, and if the value of $(g_i\ T)$ is very large compared with others, then the slight change of T's value might cause the very large change of the GP tree's output. As shown in (3), the FNS method penalizes the GP tree if the tree contains such nodes.

$$\vartheta_{EF} = \vartheta_{MSE} + 0.1\hat{\vartheta}_{MSE} + \vartheta_{FNS} \tag{3}$$

where $\vartheta_{FNS} = 0$ if for all g_i in the GP tree are

$$|(g_i\ T) - f_i(T)| \le \delta \text{, otherwise } \vartheta_{FNS} = \alpha$$

f_i is a mathematic function corresponding to g_i, α is very large positive number such as 1.0E10, and δ (0.1) is a tolerance. If $(g_i\ T)$ is large, then $(g_i\ T)$ is very different from f_i, and ϑ_{FNS} becomes α. In this case, the Hooke & Jeeves method tries to make $(g_i\ T)$ approach to f_i within the tolerance δ by estimating weights of the GP tree through minimizing ϑ_{EF}. If this effort fails, the GP tree has a very large fitness value, and will be excluded in the next generation. Note that in FNS, the only way of $(g_i\ T)$ approximating f_i is that T has a small value because g_i is a low order polynomial.

2.3 Managing the polynomials

During the evolving process, many GP trees produce too complex polynomials although low order Taylor series are used. There is need for reducing the complexity of a polynomial. One way of doing this is to give the allowable maximum number of term($n_{\max}$) of polynomials to the GP system so as not to generate GP trees representing a too complex polynomial. The algorithm for computing $n_{\max}$ of polynomial from the GP tree can be implemented using the stack structure. Its description is rather lengthy. So, we will not discuss due to the space limitation.

Table 1 Parameters used in GP.

Max. generation	40
Selection method	Tournment with 30 trees
Max. terms of a polynomial	100
Reproduction probability	0.15
Crossover probability	0.7

3 Test for the GP as a Data Miner

The function below shows the Goldstein-price function typically used as a benchmark problem for testing the performance of the optimization algorithms.

$$f(x_1,x_2)=(1+(x_1+x_2+1)^2\cdot(19-14x_1+3x_1^2-14x_2+6x_1x_2+3x_2^2))$$
$$\cdot(30+(2x_1-3x_2)^2\cdot(18-32x_1+12x_1^2+48x_2-36x_1x_2+27x_2^2))$$

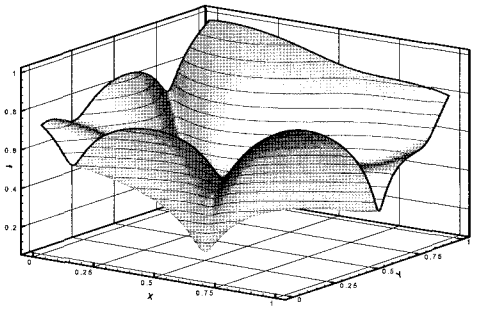
$$-2\le x_i\le 2 \qquad i=1,2$$

It is nearly impossible to approximate the Goldstein-price function with a good accuracy because the function value is changed from 0 to 1.0E6, and this large value is too dominant to others. So, we take the logarithm scale($f_{\log}(x_1,x_2)=\log(1+f(x_1,x_2))$).

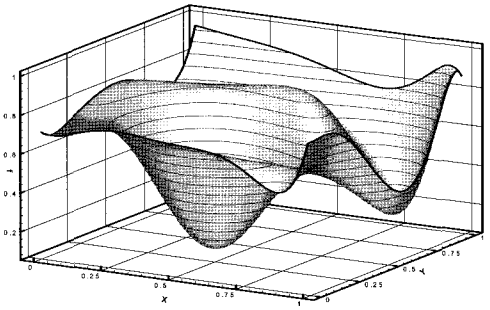
The learning set is prepared as only 5x5 grid type, and in the same manner the test set is made by 200x200 grid type. The size of the population is 5000. The learning and test set are normalized. Figure 1 (a) shows $f_{\log}(x_1,x_2)$. Since this function is highly nonlinear and the size of the learning set is only 25, do not expect for GP to produce the good GP tree. Figure 1 (b) shows the results of GP when (1) is used for the fitness function. The GP tree is severely overfitted. On the other hand, as shown in Figure 1 (c), when (3) used as the

fitness function, the results give the smooth surface, and roughly picture the overall feature of $f_{\log}(x_1, x_2)$.

Using the translation program, the GP tree is transformed into the normal polynomial form, and this polynomial is simplified by Mathematica. The result is given in 2. The size of polynomial is manageable.

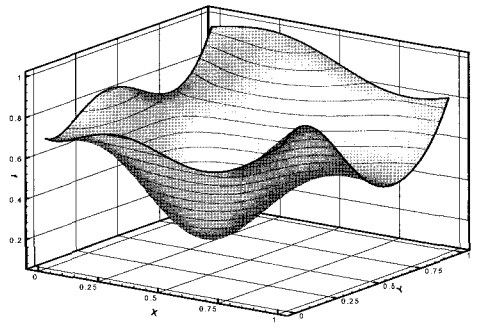


(a) Original function.



(b) Results of GP using (1) as a fitness function.

The MSE of learning and test set are 3.386E-6 and 0.0712, respectively.



(c) Results of GP using (3) as a fitness function.
The MSE of learning and test set are 3.645E-3 and 4.365E-3, respectively.

Figure 1: The logarithm-scaled Goldstein-price function.

```
0.693707 (-1.94697 (-2.56609 + 4.154 x2)
    (1. - 0.253758 (1.76982 - 3.07988 x1 - 1.45075 x2)^2
      (1. - 0.520818 x2^2)^2)
    (0.598926 (1. + 0.556292 (1.66548 x1 + 0.751999
          (-1.0079 x1 + 0.892572 x2)^3 - 1.47491 x2) ^ 2) -
      0.97981 (1. - 0.439762 (-1.20691 x1^3 + x1^2 (3.11693 -
          2.61313 x2) - 1.88593 x1 (-2.20564 + x2)
          (-0.179953 + x2) - 0.4537 (-2.94709 + x2)
          (-1.1928 + x2) (0.561499 + x2)) ^2)) +
  1.542 (1.14388 (-1.08873 + x2) (0.0949399 + x2) +
    1.28987 (-0.327209
        (1. - 0.0880898 (1.66792 - 4.38709 x1 - 0.462422 x2)^2
          (1. - 0.0417391 (1.10438 x1 - 0.286908 x2)^2
            (-2.54243 + x2)^2 (1.16337 + x2)^2) ^2) +
      0.846979 (1. + 0.341689 (1. - 0.292192
          (1.64987 - 1.24615 x1 - 3.537 x2)^2
          (1. - 0.473873 x2^2)^2) ^2))))
```

Figure 2: The polynomial obtained by GP.

4 Data Mining System for ship design

In this section, Data Mining system for data utilization by using polynomial genetic programming (PGP) and other modules is presented(Lee et al 2006). That is, the tool is contrived to apply to ship design under the case that the accumulated data is not enough to make learning process. The system is applied to real ship design problem in a Korean shipyard. Figure 3 shows the developed Data Miner by using GP. The DM system can make fitting functions with 3 types of GP such as GP with high order polynomial, linear model GP with polynomial (PLM-GP), and linear model GP with math functions (LM-GP). Users can make the process of function approximation by selecting arbitrary functions that they want to use. Also designer can set the number of polynomial terms and maximum allowable order of polynomial. And the generated function tree can be converted to C code in order to integrate with other program as shown in Figure 4. The system is implemented by using Microsoft Visual Studio .Net C# programming. In order to adapt the developed system in real ship design problem, model test for the estimation of the performance of propulsion system (K_T) is applied. Unfortunately, the experiment data for the K_T is secret. So we generate 1000 data by using empirical formula. It means that the data contains some noisy. The system automatically uses 800 data for training, and 200 data for test among 1000 learning data. Figure 5 shows the test result for 200 test data. The solid line means the hoped location of target values. The dotted points are estimated results. The reasonable range of K_T is between 0.15 and 0.35, in which this system can estimate the K_T value very well.

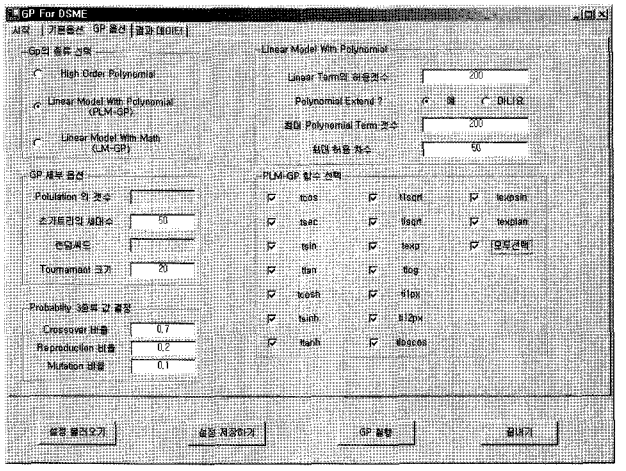


Figure 3: Data Miner for ship design by using GP

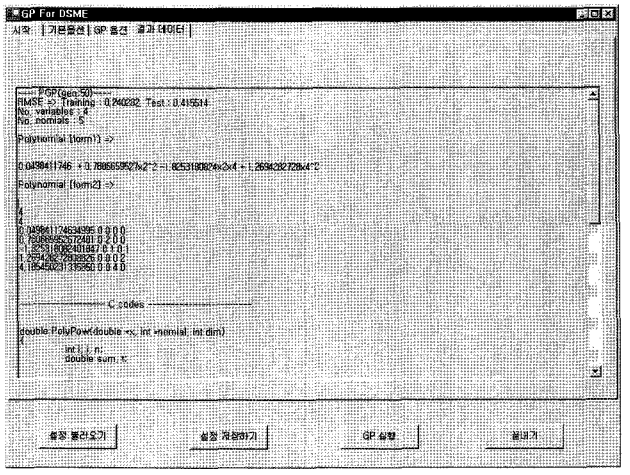


Figure 4: C code generation for an approximated function by GP tree in Data Miner

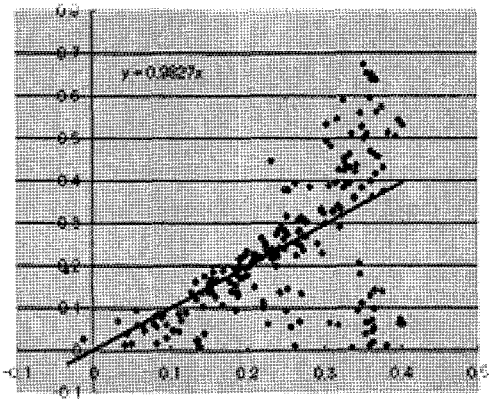


Figure 5: Estimated value for K_T

5 Conclusions

Generally, accumulated existing ship data is very valuable, in which all sorts of design knowledge and yard practices are melted. Recently the importance of the utilization of engineering data is gradually increasing. In this paper, Data Mining system as a data approximation/prediction tool to assist the ship designing process with insufficient learning samples is developed.

First of all, Polynomial genetic programming technique is presented to make approximated function from the accumulated existing data. And also the Data Miner as a data miming tool based on the PGP is presented. The system can give consistent results with limited amount of learning samples, regardless of whether or not samples contain noise. The validation test and the adoption of the developed method in the ship designing process showed that the method is good for non-linear function approximation with limited amount of learning data, without overfitting.

The Data Mining system is used in real preliminary ship design process to generate new empirical formula for new concept ships, such as LNG carrier, FPSO, and so on.

Acknowledgements

This work is supported by Advanced Ship Engineering Research Center (R11-2002-104-08002-0).

References

Alotto, P., M. Gaggero, G. Molinari and M. Nervi. 1997. A Design of Experiment and Statistical Approach to Enhance the Generalized Response Surface Method in the Optimization of Multi-Minimas, IEEE Transactions on Magnetics, **33, 2**, 1896-1899.

Ishikawa, T. and M. Matsunami. 1997. An Optimization Method Based on Radial Basis Function, IEEE Transactions on Magnetics, **33, 2/II**, 1868-1871.

Koza, J.R. 1992. Genetic Programming: On the Programming of Computers by Means of Natural Selection, The MIT Press.

Lee, K.H., Y.S. yeun, J.H. Lee and J. Oh. 2006. Data Analysis and Utilization Method Based on Genetic Programming in Ship Design, Lecture Notes in Computer Science, 3981.

Malik, Z., H. Su and J. Nelder. 1986. Informative Experimental Design for Electronic Circuits, Quality and Reliability Engineering, **14**, 177-188.

Myers, R.H. and D.C. Montgomery. 1995. Response Surface Methodology: Process and Product Optimization Using Designed Experiments, John Wiley & Sons, Inc.

Ott, R.L. 1993. An Introduction to Statistical Methods and Data Analysis, Wadsworth Inc.

Simpson, T.W., J.K. Allen and F. Mistree. 1998. Spatial Correlation and Metamodels for Global Approximation in Structural Design Optimization, Proc. of DETC98, ASME.

Yeun, Y.S., K.H. Lee and Y.S. Yang. 1999. Function Approximations by Coupling Neural Networks and Genetic Programming Trees with Oblique Design Trees, AI in Engineering, **13, 3**.