MS2006
Proceedings of the International Conference on Modeling and Simulation 2006
28 – 30 August 2006, Konya, TURKEY
Paper No:B070

USING MACHINE LEARNING METHODS IN THE COMPUTATION OF ENGINE POWER OF MOTORBOATS AND THE COMPARISON OF THESE METHODS

Bal,M.[1],Amasyalı,M.F.[2], Çelebi,U.B.[3], Ekinci,S.[3]
[1] Faculty of Chemistry and Metallurgical, Mathematical Engineering Department
[2] Faculty of Electrics and Electronics, Computer Engineering Department
[3] Faculty of Mechanical Engineering, Naval Architecture and Marine Engineering Department
Yıldız Technical University, Istanbul, TURKEY
mertbal@ yildiz.edu.tr, mfatih@ce.yildiz.edu.tr,
ucelebi@ yildiz.edu.tr, ekinci@yildiz.edu.tr

ABSTRACT
Determining main engine power has importance in developing the entire design parameters of a motorboat. After developing motorboat hull from the principal dimensions, all of the resistance forces must be evaluated. However, it is extremely complicated to determine engine power both overcoming these forces and maintaining the required speed mentioned in the specifications prepared. It is possible to say that determining principal parameters by machine learning methods is superior to traditional computation techniques for time and cost. In this study, 8 different machine learning techniques are applied for determining the engine power based on 5 different combinations of principal parameters of current motorboats. The principal parameters are ship length, breadth, draught and speed. The results of these experiments are compared to examine which method and feature set give better results. Pace Regression and Model Trees are over performed than others on our motorboat datasets.

Keywords: Machine Learning, Neural networks, Regression, Engine Power Prediction, Motorboat

INTRODUCTION
For a ship to sail in safe, resistance forces by air and water must be calculated and an engine should be selected which will produce enough power output to overcome these forces. Therefore, a minimal trust should be applied to the ship which cannot be less than the resistance forces obscuring the motion of the ship. Since this trust is equal to the resistance force, ship resistance is used in common instead of the trust. For this reason, power of the main engine, the trust maintained at the end of the ship after the transmission losses from main engine to propeller and the propeller losses must meet the total resistance of the ship. Besides, the ship must reach the required speed with the engine chosen.

At the predesign stage, the necessary information to calculate ship resistance or main engine power is usually obtained by testing similar ship models in the towing tank. These data are converted into characteristic curves, tables and empiric equations via principal dimensions of ship. Moreover, with the developing computer technology, today it is also possible to calculate the ship resistance and power using specific computer programs for the dimensions of the ship as inputs.

Recently, calculating design parameters with neural network is found to be much better than the traditional calculation methods in terms of time and costs. In the literature there are some applications of neural networks into naval architecture: neural network applications in naval architecture and marine engineering [1], design of a robust neural network structure for determining initial stability particulars of fishing vessels [2], modeling and simulation of ship main engine based on neural network [3], determination of approximate main engine power for chemical cargo ships using radial basis function neural network [4] can be given as examples.

In this study, different machine learning regression techniques are applied for determining the engine power based on principal parameters of current motorboats which have a length of 8 to 25 meters.

The remainder of the paper is organized as follows: In the next section, a motorboat's main dimensions are showed. Brief descriptions of used machine learning (ML) methods are mentioned in the third section. The results obtained from different ML methods and feature sets are given at the fourth section. In the last section conclusions are summarized.

A Motorboat's Main Dimensions and Related Definitions

A motorboat's profile and upper view are given below in Figure 1. Main dimensions and related parameters in the illustrations are defined below.

Fig.1. A motorboat's profile and upper view

(BL) Baseline: The horizontal line parallel to the design waterline (DWL), which cuts the midship section at the lowest point of ship. The vertical heights are usually measured from the baseline.
(L=L_{OA}) Length Overall: The total length of the ship from one end to the other, including bow and stern overhangs.
(T) Draught: The vertical distance from the waterline at any point on the hull to the bottom of the ship.
(B) Breadth: The distance from the inside of plating on one side to a similar point on the other side measured at the broadest part of the ship.
(V) Ship speed (Knot): The distance in miles taken in an hour.
(F_r) Froude Number: It is the dimensionless parameter which is calculated by Equation 1.

$$F_r = \frac{V}{\sqrt{gL}} \tag{1}$$

In Equation 1, V is the ship speed (m/s), L is the length of a ship (m) and g is the gravitational acceleration (m/s^2).

MACHINE LEARNING METHODS
In this work, we used WEKA's (available at www.cs.waikato.ac.nz/ml/weka) 8 different regression algorithms. All experiments are done with WEKA' s default parameters.

Least Median Square
Least median square algorithms use the estimates of the gradient vector from the available data. LMS incorporates an iterative procedure that makes successive corrections to the weight vector in the direction of the negative of the gradient vector which eventually leads to the minimum median square error. [5]

Lineer Regression
Models will be formed by random sub samples in linear regression and the one that is convenient to Akaike criterion [6] will be selected.
A linear regression equation can be specified as:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{2}$$

The driving idea behind the AIC is to examine the complexity of the model together with goodness of its fit to the sample data, and to produce a measure which balances between the two.
Its formula is

$$AIC = 2k - 2\ln(L) \tag{3}$$

where k is the number of parameters, and L is the likelihood function. Usually however, normally distributed errors are assumed and AIC is computed as

$$AIC = 2k + n\ln(RSS/n) \tag{4}$$

where n is the number of observations and RSS is the residual sum of squares.

Multi Layer Perceptron
In Multi Layer Perceptrons, backpropagation is used generally. The backpropagation algorithm was derived from a Least Mean Squares Approach and a description of this may be found in Rumelhart and McClelland [7]. What follows is a description of how to implement the algorithm which is essentially derived from Lippmann [8] If the bias terms are imagined as an input which is always set to 1 and fed through a weight, The output equation of basic element of the network is given in Equation 5.

$$Output_j = S(\sum_{i=0}^{N} w_{ij} x_i) \tag{5}$$

where S is a sigmoid non-linearity activation function(Equation 6).

$$S(a) = \frac{1}{1 + Ae^{-a}} \tag{6}$$

where A determines the level of non-linearity, x_i is input i and w_{ij} is the weight connecting input i to neuron j.

First the error for the output layer nodes is computed by using Equation 7.

$$E_j = (t_j - a_j) a_j (1 - a_j) \tag{7}$$

where E_j is error for node j of the output layer, t_j is target activation for node j of the output layer and a_j is actual activation for node i of the output layer.

Then, successively, the error values for all hidden layer nodes are computed by using Equation 8.

$$E_i = a_j (1 - a_j) \sum_j E_j w_{ij} \tag{8}$$

where E_i is error for node i in a hidden layer, E_j is error for node j in the layer above, w_{ij} is weight for the connection between node i in the hidden layer and node j in the previous layer and a_i is activation of node i in the hidden layer.

At the end of the error backward propagation phase, nodes of the network (except the input layer nodes) will have error values. The error value of a node is used to compute new weights for the connections which lead to the node. Very generally, the weight change is done by using Equation 9.

$$w_{ij} = w_{ij} + \Delta w_{ij} \tag{9}$$

where w_{ij} is weight of the connection between node i in the previous layer and node j in the output layer or in a hidden layer and Δw_{ij} is weight change for the connection between node i in the previous layer and node j in the output layer or in a hidden layer.

Pace Regression

Classical ordinary least squares (OLS) regression is simple and computationally cheap but it has generally unsatisfactory models. If there is large number of attributes, many of which are redundant, OLS models has worse performance than the models taking fewer attributes into account. There are many researches on neutralizing this effect by selecting a subset of attributes. Pace regression improves on classical ordinary least squares (OLS) regression by evaluating the effect of each variable and using a clustering analysis to improve the statistical basis for estimating their contribution to the overall regression. [9]

Radial Basis Functions

The Radial Basis Function (RBF) networks have a simple architecture comprising an input layer, a single (usually Gaussian) hidden layer and an output layer, the nodes of which are expressed as a linear combination of the outputs of the hidden layer nodes [10]. For a one-node output layer, the global input-output relationship of an RBF Neural Network can be expressed as a linear combination of K basis function as follows:

$$f(x) = \sum_{i=1}^{K} w_k \Phi_k(x) \tag{10}$$

where $x = [x_1, x_2, \ldots\ldots\ldots\ldots, x_M]^T$ is the M-dimensional input vector, w_k are the weighting coefficients of the linear combination, and $\Phi_k(x)$ represents the response of the k^{th} neuron of the hidden layer. Typically, the basis function $\Phi_k(x)$ are assumed to be Gaussian shaped with scale factor σ_k; their values decrease monotonically with the distance between the input vector x and the center of each function $c_k = [c_{k1}, c_{k2}, \ldots\ldots\ldots\ldots, c_{km}]^T$

$$\Phi_k(x) = \exp\left(-\frac{\|x - c_k\|^2}{\sigma^2}\right) \tag{11}$$

Support Vector Regression

The basic idea in Support Vector Regression (SVR) is to map the input data x into a higher dimensional feature space F via a nonlinear mapping Φ and then a linear regression problem is obtained and solved in this feature space [11]. Therefore, the regression approximation addresses the problem of estimating a function based on a given data set $G = \{(x_i, d_i)\}_{i=1}^{l}$ (x_i is input vector, d_i is the desired value). In Support Vector Machines (SVM) method, the regression function is approximated by the following function:

$$y = \sum_{i=1}^{l} w_i \Phi_i(x) + b \tag{12}$$

where $\{\Phi_i(x)\}_{i=1}^{l}$ are the features of inputs $\{w_i\}_{i=1}^{l}$ and b are coefficients. The coefficients are estimated by minimizing the regularized risk function

$$R(C) = C\frac{1}{l}\sum_{i=1}^{l} L_\varepsilon(d_i, y_i) + \frac{1}{2}\|w\|^2 \tag{13}$$

where

If $|d - y| \geq \varepsilon$ then $L_\varepsilon(d, y) = |d - y| - \varepsilon$, otherwise $L_\varepsilon(d, y) = 0$

and ε is a prescribed parameter.

Simple Linear Regression

It is the process of fitting straight lines between each attribute and output. In Equation 14, the values of w and w_0 are estimated by the method of least squares.

$$y = wx + w_0 \tag{14}$$

The line (model) having lowest squared error is selected as the final model among each attributes model [12].

M5 Model Trees

M5 algorithm uses the following idea: split the parameter space into areas (subspaces) and build in each of them a local specialized linear regression model [13]. The splitting in Model Tree (MT) follow the idea used in building a decision tree, but instead of the class labels it has linear regression functions at leaves, which can predict continuous numeric attributes. Model trees generalize the concepts of regression trees, which have constant values at their leaves. So, they are analogous to piecewise linear functions (and hence non-linear). Model trees learn efficiently and can tackle tasks with very high dimensionality – up to hundreds of attributes. The major advantage of model trees over regression trees is that model trees are much smaller than regression trees, the decision strength is clear, and the regression functions do not normally involve many variables.

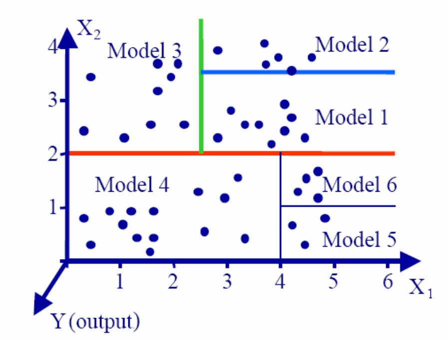

Fig 2. Splitting the input space $X_1 \times X_2$ by M5 model tree algorithm; each model is a linear regression model $y = a_0 + a_1x_1 + a_2x_2$.

EXPERIMENTAL RESULTS

In this study, engine power is modeled by different ML algorithms and different input sets. Our problem can be seen as a regression problem since the predicted variable is continuous. The regression algorithm names and abbreviations are given at Table 1.

Table 1. Regression Methods and Abbreviations

Method Name	Abbreviation
LeastMedSq	LMS
LinearRegression	LR
Multi Layer Perceptron	MLP
PaceRegression	PR
RBF	RBF
Support Vector Regression	SVM
SimpleLinearRegression	SLR
Model Trees	M5P

We used 5 different feature set as inputs. They are obtained from combination and processing of 4 principal dimensions of a motorboat. The first feature set (A) consists of the common parameters using at a motorboat's design. The second set (B) is obtained from first set by a WEKA's feature selection method (cfsSubsetEval). The third set is the main components of first set. The forth and fifth sets are calculated from the first and third sets by projecting them onto their best three principal components. All input sets and abbreviations are given at Table 2.

Table 2. Used Feature Sets and Abbreviations

Feature Set	Abbreviation
L-B-T-V-L/B-B/T-F_r	A
L	B
L-B-T-V	C
3 features (PCA on A)	D
3 features (PCA on C)	E

For generalization of results, 5 fold cross validation is applied. For each fold, training and test sets are composed of 202 and 48 samples respectively.

In Figure 3, MLP's predictions are given with single inputs. X and Y axis show the single input and the output of the functions respectively. Mean Square Error (MSE)'s are given at the top of each experiment. It can be seen that L is the best parameter.

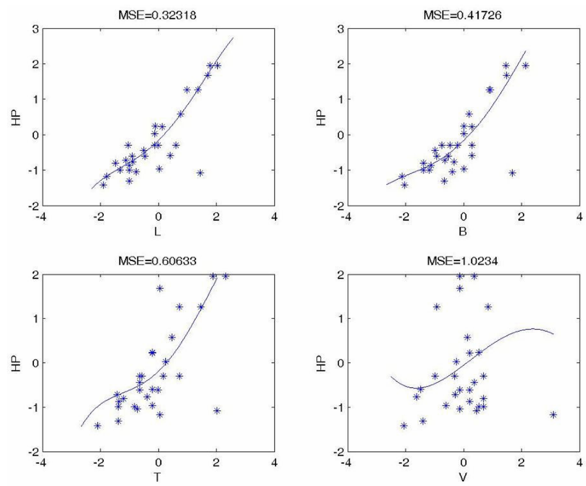

Fig.3. HP predictions using single inputs with MLP

In Table 3, engine power prediction experiments are summarized. The values are the means of 5 fold experiment's MSEs. For the explanations of methods and feature set's abbreviations see Table 1 and 2. The last column gives the mean errors of methods over all feature sets. The last row gives the mean errors of feature sets obtained from all ML methods.

Table 3. Prediction Errors

	Feature Sets					
Methods	A	B	C	D	E	Mean
LMS	0,1986	0,2023	0,1681	0,1884	0,1693	0,1854
LR	0,1944	0,2026	0,159	0,181	0,1743	0,1822
MLP	0,1836	0,2588	0,1812	0,1888	0,1887	0,2002
PR	0,1217	0,2026	0,1586	0,1812	0,1742	0,1677
RBF	0,863	0,5929	0,7704	0,8012	0,8118	0,7679
SVM	0,1683	0,2018	0,1678	0,1805	0,1735	0,1783
SLR	0,2026	0,2026	0,2026	0,3305	0,2685	0,2414
M5P	0,125	0,1995	0,1371	0,1399	0,1651	**0,1533**
Mean	0,2572	0,2578	0,2431	0,2739	0,2657	

CONCLUSIONS

In naval engineering, in order to calculate the engine power and the water displacement, classical methods use several parameters [14], some of which require tedious work and long experiments. This study aims to estimate the engine power and displacement by the help of only easily accessible parameters. As estimation models, several machine learning techniques are used. In this study, two analyses have been conducted to find:

- The most effective input(s)/attribute(s) to determine the output
- The most successful ML algorithm

For the first one, five different training and test sets have been created by using different feature sets. For the second analysis, 8 ML algorithms have been used.

Some important results have been summarized as follows:

- The best/lowest two error values are obtained with the A feature set. But the C feature set (including all principal parameters of the motorboats) has mean lowest error rate.
- There is no significant difference between mean error rates of the feature sets while ML methods differ.
- The lowest error rate is 0.12. It is obtained with Pace regression algorithm on the A feature set. The constructed Engine Power (EP) linear model is given above.

$$EP = 9.39 + 0.78L - 2.44B + 0.15V - 3.42L/B - 1.44FR$$

- RBF has the worst/highest mean error rate.
- M5P is the best performed algorithm over all feature sets.
- While M5P have the best errors on four out of five feature sets, PR has the best error on one set.
- SVM, LMS, LR and MLP have lowest errors with C set, PR and M5P with A set, RBF with B set and SLR with A,B,C sets.

As a result, Pace Regression and Model Trees algorithms are suitable for estimation of engine power of motorboats. They achieve best performances with A feature set.

REFERENCES

1. Raya, T., Gokarna, R.P., Shaa, O.P., Neural network applications in naval architecture and marine engineering. Artificial Intelligence in Engineering 10 (3), 213–226, 1996.
2. Alkan A.D., Gulez K., Yilmaz H., Design of a robust neural network structure for determining initial stability particulars of fishing vessels, Ocean Engineering Vol.31, 761-777,2004.
3. Xiao J.,Wong X., Boa M., The modeling and simulation of ship main engine based on Neural Network, Shanghai Maritime University, Proceedings of the 4th World Congress on Intelligent Control and Automation, 2002.
4. Kapanoğlu B.,Çelebi U.B.,Ekinci S.,T.Yıldırım T., Determination Of Approximate Main Engine Power For Chemical Cargo Ships Using Radial Basis Function Neural Network, Turkish Naval Academy, Journal of Naval Science and Engineering, Volume 2, Number 2, pp:115-115, 2004.
5. Peter J. Rousseeuw, Annick M. Leroy, Robust regression and outlier detection, 1987.
6. http://en.wikipedia.org/wiki/Akaike_Information_Criterion.
7. Rumelhart, D.E. McClelland,J.L., Parallel Distributed Processing:Explorations in the Microstructure of Cognition, MIT Press, 1986.
8. Lippmann,R.P., An Introduction to Computing with Neural Nets, IEEE ASSP Magazine, 1987.
9. Wang, Y., A new approach to fitting linear models in high dimensional spaces, PhD Thesis. Department of Computer Science, University of Waikato, New Zealand, 2000.
10. Corsini G., Diani M., Grasso R., De Martino M., Mantero P., Serpico S.B. ,Radial Basis Function and Multilayer Perceptron Neural Networks for Sea Water Optically Active Parameter Estimation in Case 2 Waters: A Comparison, International Journal Remote Sensing 24(20) ,2003.
11. Wang, W., Xu, Z., A Heuristic Training for Support Vector Regression, Neurocomputing, vol: 61, pp. 259-275, 2004.
12. www.oxfordjournals.org/tropej/online/ma_chap2.pdf.
13. Bhattacharya, B., Solomanite, D.P., Neural Network and M5 Model Trees in Modelling Water Level-Discharge Relationship, Neurocomputing, vol :53, pp. 381-396, 2005.
14. Baykal R., Dikili A.C., Ship Resistance and Engine Power, I.T.U, ISBN 975-561-196-7, Istanbul (in Turkish),2002.