대한조선학회논문집
Journal of the Society of Naval Architects of Korea
eISSN:2287-7355, Vol. 57, No. 6, pp. 312-321, December 2020
https://doi.org/10.3744/SNAK.2020.57.6.312

선형변수 기계학습 기법을 활용한 저속비대선의 잉여저항계수 추정

김유철[1]·양경규[2]·김명수[1]·이영연[1]·김광수[1,†]
선박해양플랜트 연구소[1]
충남대학교 선박해양공학과[2]

Prediction of Residual Resistance Coefficient of Low-Speed Full Ships Using Hull Form Variables and Machine Learning Approaches

Yoo-Chul Kim[1]·Kyung-Kyu Yang[2]·Myung-Soo Kim[1]·Young-Yeon Lee[1]·Kwang-Soo Kim[1,†]
Korea Research Institute of Ships and Ocean Engineering (KRISO)[1]
Department of Naval Architecture and Ocean Engineering, Chungnam National University[2]

In this study, machine learning techniques were applied to predict the residual resistance coefficient (Cr) of low-speed full ships. The used machine learning methods are Ridge regression, support vector regression, random forest, neural network and their ensemble model. 19 hull form variables were used as input variables for machine learning methods. The hull form variables and Cr data obtained from 139 hull forms of KRISO database were used in analysis. 80 % of the total data were used as training models and the rest as validation. Some non-linear models showed the overfitted results and the ensemble model showed better results than others.

Keywords : Machine learning(기계학습), Cr prediction(잉여저항계수 추정), Low-speed full ship(저속비대선), Hull form variables(선형변수), Regression(회귀 분석)

1. 서 론

선박의 설계 단계에서 선형의 성능을 예측하기 위한 방법으로 모형시험과 CFD (Computational Fluid Dynamics)가 주로 이용되고 있다. 통상 몇 개의 후보 선형(최근에는 상당히 많은 후보 선형에 대하여 CFD 계산을 활용하는 경우가 늘고 있음)에 대하여 lines를 생성하고, CFD 계산을 통해서 최적 선형을 선정한 후, 모형시험을 통하여 최종 성능을 확인하는 단계를 거치며 설계가 진행된다. 하지만 lines를 결정하기 이전의 단계에서 주요 제원을 선정하기 위해 선박의 성능을 예측하기 위한 방법으로서 모형시험과 CFD를 활용하는 것은 불가능하다. 이러한 초기 설계 단계에서는 통계적 예측 방법이 여전히 사용되고 있다.

계열 시험 자료를 사용하는 방법과 모형 시험 결과를 통계적으로 분석하는 방법으로 대변되는 통계적 예측 방법은 1930년대 테일러 표준 계열 (Taylor standard series; Taylor, 1933) 실험을 시작으로 Gertler 차트 (Gertler, 1954), Lap 차트 (Lap, 1956), Guldhammer & Harvald (1965)의 실험 데이터 등이 발표되었으며, Holtrop의 방법 (Holtrop and Mennen, 1978; Holtrop and Mennen, 1982; Holtrop, 1984)이 가장 잘 알려져 있고 현재에도 약간의 계수 보정을 통해서 사용되고 있다. 근래에 단순 신경망 이론을 접목하여 선박의 저항 및 추진 성능을 예측하는 방법이 Kim and Park (2015)에 의해 시도된 바 있으나, 한정된 조선소의 설계 선형들에 대한 분석으로 데이터의 다양성이 다소 떨어지는 단점을 갖고 있다. Kim et al. (2019)은 비교적 최근의 KRISO (Korea Research Institute of Ships and Ocean Engineering) 모형시험 데이터를 회귀 분석하여 저속비대선의 잉여저항계수 추정 회귀식을 소개하고, 이를 활용하여 선형 설계를 수행한 결과를 발표한 바 있다. Cho et al. (2019)은 KVLCC2를 기본 선형으로 오프셋 변환을 수행한 1,263척의 선형데이터를 생성하고, 이들 3차원 좌표와 Holtrop (1984)의 식으로 예측한 저항 결과를 학습시킨 결과를 최근에 소개하였다.

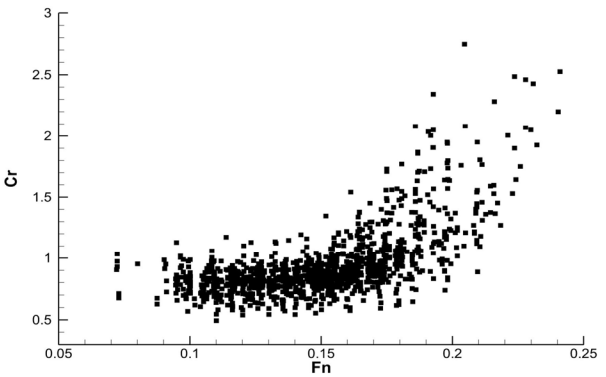
최근 컴퓨팅 능력의 비약적 발전과 병렬 컴퓨팅 시스템의 보

Received : 21 May 2020 | Revised : 23 July 2020 | Accepted : 10 August 2020
† Corresponding author : Kwang Soo Kim, ksookim@kriso.re.kr

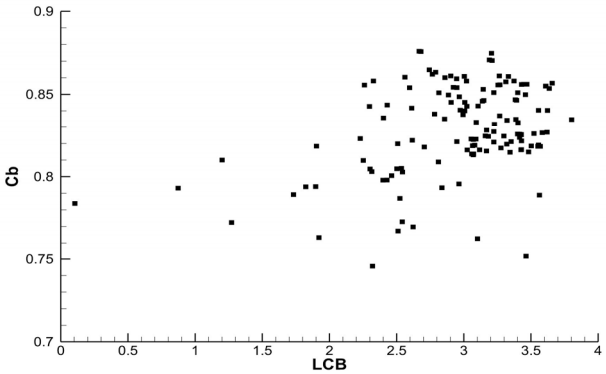
급에 따라 선형 설계에 있어서 CFD의 역할 (Kim et al., 2007; Kim et al., 2011; Park et al., 2014; Park et al. 2015)이 상대적으로 높아진 상태이나, 이와 더불어 기계학습 및 딥러닝으로 대표되는 데이터 분석 기법 역시 비약적인 발전으로 양질의 많은 데이터가 주어지면 그동안 생각하지 못했던 분석 결과들을 주는 사례 역시 증가하고 있는 시대이다. 본 연구는 선형 변수, lines 정보, 실험 결과를 바탕으로 기계 학습을 통하여 선형 설계에 도움이 되는 데이터 분석의 초기 단계로서 기존의 선형 회귀식을 대체할 수 있는 다양한 기계학습 모델을 검토하고, 이를 이용한 선형 설계 결과를 보인다. KRISO가 보유한 저속비대선의 저항 시험 결과와 DB화 된 선형 변수를 연결하는 몇 가지 기계학습 모델에 대한 적용을 통하여 각각의 결과 및 특성을 파악하고, 이들을 조합한 최적의 모델을 결정한다. 도출된 모델의 검증을 위하여 Supramax 선형과 이를 기초로 설계된 파생 선형들의 잉여저항성분을 기계학습 모델, 선형 회귀, CFD 계산을 이용하여 예측하고 비교를 수행한다.

2. 분석 대상 및 선형 변수

본 연구에서는 국내 건조 선형 중 상당 부분을 차지하고 있는 탱커선과 벌크선을 포함하는 저속비대선의 잉여저항 계수를 분석 대상으로 선정하였다. 분석에 사용된 선형의 수는 이상치(outlier)를 제외한 총 139척으로 모든 선형에 대하여 기본 제원 이외에 선형의 국부 형상을 나타낼 수 있는 상세 선형 변수를 추출하였다. 선형의 용적 분포는 MR급(Medium Range)에서 VLCC급(Very Large Crude Carrier)까지 포함한다. 흘수 조건은 설계 흘수 조건만을 고려하였다. Fig. 1은 본 연구의 대상 선에 대한 Cr(Coefficient of residual resistance; 잉여저항계수) 분포와 LCB-Cb 분포를 보인다. 저속비대선의 특성상 LCB(종방향 부력 중심)가 선수 쪽으로 치우친 경향을 보이며 Cb(방형계수)는 0.8~0.85 부근에 집중되어 있는 것을 확인할 수 있다.



(a) Fn-Cr plot



(b) LCB-Cb plot

Fig. 1 Hullform data distribution of target hullforms

잉여저항계수 추정시 선형의 전체적인 형상 이외에 국부적인 선형 변화를 일부 반영하기 위하여 선형의 국부적 특성들을 나타낼 수 있는 선형 변수들을 정의하고 분석에 사용하였다. 선형 변수는 Kim et al. (2019)에 소개된 바와 같이 주요 제원과 형상 계수들, 유체정역학적 계수들, Cp 곡선과 관련된 변수들, 선수, 선미 형상 관련 계수들로 구성되어 있다. 본 연구에서는 기계학습 모델의 성능 비교를 위하여 Kim et al. (2019)이 선형 회귀식(Linear regression) 구성에 사용한 선형 변수들 혹은 이와 유사한 변수들을 선정하여 분석에 사용하였다.

3. 기계학습 기법의 적용

3.1 데이터 준비 및 분류

Kim et al. (2019)은 회귀식 도출을 위하여 먼저 잉여저항계수를 Fn(프루드 수)의 다항식으로 근사하여 하나의 변수로 사용하였다. 그리고 선형 변수들을 다항식으로 근사한 변수에 곱하거나 더해서 회귀식을 구성하였다. 회귀식에 사용된 변수들은 Table 1과 같다. 분석 대상인 저속비대선을 Cb에 따라서 상대적으로 비대형 선형과 세장형 선형으로 나누고 특정 Fn를 기준으로 저속 구간과 고속 구간을 나누어 총 4개의 회귀식을 도출한 바 있다. 이렇게 도출한 회귀식은 전체 데이터에 대한 결정계수(R^2)가 약 0.91로 계산되었으며, 이를 전저항(Rt)으로 환산했을 경우 약 1.7 %의 오차를 보였다.

본 연구에서는 저속비대선의 잉여저항계수 회귀식을 도출한 Kim et al. (2019)의 결과와 비교를 수행한다. 따라서 기계학습에 사용한 변수는 Kim et al. (2019)이 사용한 변수와 유사하게 사용하였으며, 일부 변수를 대체하기 위한 특정 스테이션에서의 Cp값을 추가하였다. Table 2는 기계학습에 사용된 변수를 보인다. Kim et al. (2019)의 회귀식에 사용된 변수 대부분을 그대로 사용하고 있으며, τ대신 2.5 스테이션에서의 폭을 선폭으로 나눈 값을, V_{bulb} 대신 정의가 간편한 벌브 길이(L_{bulb})를, L1과 L2를 대신하기 위하여 0.7Lbp에서 선수부까지의 Cp 분포와 Fn의 제곱 항을 사용하기로 한다. 분석에 사용된 기계 학습 방법으로는 릿지 회귀, 서포트 벡터 머신, 랜덤 포레스트, 신경망 이론과 이들을 혼합한 앙상블(ensemble) 방법이며, 파이썬(Python)과 싸이킷런(Scikit-learn) 모듈을 사용하였다. 3.2부터 각 모델에 대한 개요와 적용 결과를 보이며, 모델에 대한 상세한 설명은 본 논문의 범위를 벗어나므로 생략한다.

Table 1 Local hullform variables adopted in the regression analysis (Kim et al. 2019)

Variables	Description
$L/\nabla^{1/3}$	Length-displacement ratio
RA	Run angle at station 3
EA	Entrance angle at station 19.25
Cbf	Block coeff. of fore ship
Cba	Block coeff. of aft ship
Cwf	Waterplane area coeff. of fore ship
Cpaf	Prismatic coeff. ratio of aft and fore ship
	Cpa/Cpf
τ	Parameter of shape of skeg in aft ship
	engine room breadth at 2.5 station/shaft center height
V_{bulb}	Parameter related with bow bulb volume
	bulbareaf • bulbareas • T
L1	Length between max. curvature position (Cp curve) and λ
	|λ-curvature max pos.|/ LBP
L2	λ/2LBP
λ	LWL x $2\pi Fn^2$

Table 2 Variables used in machine learning analysis

Variables
$L/\nabla^{1/3}$•crpoly*
Cpaf•crpoly*
Cb•crpoly*
L_{bulb}•crpoly*
EA•crpoly*
Cwf•crpoly*
Cbf•crpoly*
RA
Cba
Engb/Lbp
Fn^2
Cp at 0.7Lbp (cp14)
Cp at 0.75Lbp (cp15)
Cp at 0.8Lbp (cp16)
Cp at 0.85Lbp (cp17)
Cp at 0.9Lbp (cp18)
Cp at 0.925Lbp (cp19)
Cp at 0.95Lbp (cp20)
Cp at 0.975Lbp (cp21)

* crpoly: Cr obtained from Fn polynomial

데이터 분석을 수행하기 전에 독립변수에 대한 데이터 표준화를 수행한다. 서로 다른 기준이나 척도를 가진 변수를 그대로 분석에 사용하면 변수의 변화에 따른 민감도 편차가 달라져 예측이 불안정해 질 수 있으며, 많은 기계 학습 모델에서 사용되는 경사하강법(Gradient descent)에서는 표준화를 사용하면 학습이 더 쉬워지는 효과를 갖는다. 데이터 표준화는 모든 변수들을 평균이 0이고 표준편차가 1인 데이터로 바꾸는 작업으로 식 (1)로 정의한다.

$$x_{std}^{i} = \frac{x^{i} - \mu_x}{\sigma_x} \tag{1}$$

여기서 μ_x는 x의 평균값을, σ_x는 표준편차를 의미한다. 본 연구에서는 Table 2의 변수들을 식 (1)을 이용하여 표준화한 변수를 사용하여 분석을 수행하였다.

일반적으로 회귀 분석시 독립 변수의 수가 많아지면 분석 데이터에 대한 추정 결과가 향상되는 경향이 있다. 이는 회귀 모델이 분석 대상인 훈련 데이터에 과대적합(overfitting) 될 가능성이 높아지고, 새로운 데이터에 대한 추정 성능이 오히려 저하되는 결과를 가져올 수 있다. 이러한 과대적합을 방지하기 위해서 기계학습 모델을 훈련(train)시킬 때, 전체 데이터 세트를 모두 이용하지 않고 데이터의 일부를 도출된 모델의 평가(test)를 위해서 사용하는 방법이 이용된다. 이를 홀드아웃 교차 검증(holdout cross-validation)이라 한다(Fig. 2). 보통 기계학습 모델들은 하이퍼파라미터 튜닝이라는 모델의 최적 파라미터를 결정하는 모델 선택 단계를 거치는데, 이때 훈련용 데이터 세트를 다시 훈련 세트(train data set)와 검증 세트(validation data set)로 나누어 파라미터 선정을 수행한다.

홀드아웃 교차 검증은 훈련 세트와 검증 세트를 나누는 방법에 따라 추정 결과가 민감하게 변화할 수 있는 단점을 갖고 있다. 이러한 단점을 보완하기 위한 방법이 k-겹 교차 검증으로, 훈련용 데이터 세트를 훈련 세트와 검증 세트로 나눌 때, 중복을 허용하지 않고 k개의 세트로 훈련용 데이터 세트를 나누고, 그중 하나를 검증 세트로 사용하는 방법이다(Fig. 3). 이러한 과정을 k번 반복하여 평균적인 성능을 계산하고, 이를 기준으로 모델의 파라미터를 결정한다.

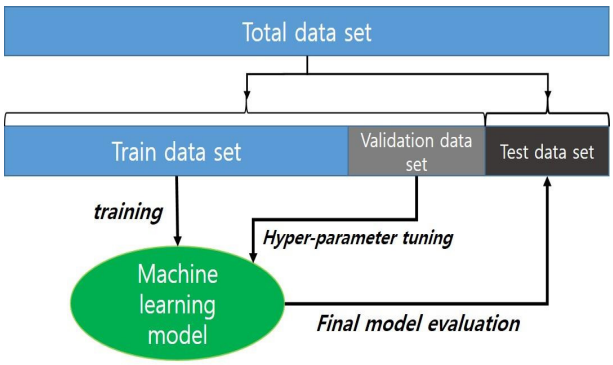


Fig. 2 Holdout cross-validation

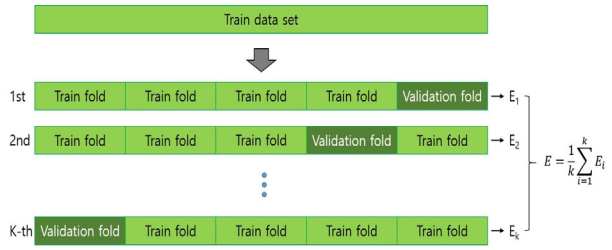


Fig. 3 K-fold cross validation

본 연구에서는 k-겹 교차 검증을 사용하여 학습을 진행하였으며, 데이터 세트를 분류할 때 개별 데이터(Cr 값 기준) 단위로 나누지 않고, 선형 단위(하나의 선형에 대한 Cr 세트)로 나누는 방식을 사용하였다(group k-fold). 이 방식을 따르면 한 가지 선형에 대한 여러 Fn의 Cr 값은 훈련 데이터 세트나 검증 데이터 세트 중 어느 한쪽에만 속하게 되어, 한 선형에 대한 데이터가 온전히 훈련 혹은 검증에만 쓰이게 된다. 본 연구에서 k는 5로 설정하였으며, 평가 세트는 총 데이터의 20 %를 무작위로 선정하여 사용하였다. 또한 Kim et al. (2019)의 연구와는 달리, 분석 대상을 Fn와 Cb로 분류하지 않고, 저속 비대선의 모든 데이터를 하나의 모델로 예측하는 방식을 사용한다.

3.2 릿지 회귀

릿지 회귀(Ridge regression)는 최소제곱법(least squares)을 사용하는 선형 회귀 방법에서 변수가 많아질 경우 생길 수 있는 과대적합을 방지하기 위하여 고안되었다 (Rifkin & Lippert, 2007). 릿지 회귀에서는 각 변수의 가중치(계수)의 절댓값을 가능한 작게 만들기 위한 규제(regularization)를 적용한다. 릿지 회귀에서는 이러한 규제의 강도를 하이퍼파라미터로 하여 튜닝을 수행하였다. 규제의 강도는 보통 스칼라 alpha로 표현하며 0에서 1 사이의 값을 갖는다. Alpha가 0인 경우는 선형 회귀와 동일한 식이 도출된다. 본 연구에서는 alpha를 [0.01, 0.05, 0.1, 0.2, 0.3, 0.5, 0.7, 1.0]으로 총 8개로 변화시키며 모델의 성능을 측정하고 최종 모델을 선정하였다.

Fig. 4는 각각의 alpha에 대하여 k-겹 검증을 수행한 결과(검증 데이터에 대한 결정계수, R^2를 점수(score)로 표기)를 보인다. 현재 데이터에 대해서는 alpha가 커질수록 검증 점수가 낮아지는 것을 확인할 수 있다. 따라서 릿지 회귀 모델의 최종 모델은 alpha를 0.01로 결정하고, 훈련용 세트 전체에 대하여 최종 학습을 진행하였다.

최종 모델을 이용하여 훈련 세트에 적용하여 잉여저항계수 추정치를 구하여 얻어진 릿지 회귀 모델의 결정계수는 약 0.807로 계산되었다. Kim et al. (2019)의 선형 회귀식에 의한 결정 계수가 약 0.91인 것에 비해 낮은 점수이지만, Kim et al. (2019)은 모든 데이터에 적합 시킨 회귀식에 대한 결과이고 현재의 결정 계수는 훈련에 사용되지 않은 평가 세트에 대한 결과이기 때문에 직접적인 비교는 큰 의미가 없을 수 있으며 현재의 점수가 더 낮은 것은 당연한 결과로 판단된다.

Fig. 5는 잉여저항계수 예측 결과 비교를 보인다. 가로축은 회귀 결과, 세로축은 실험 결과를 나타내며, 회귀 결과가 실험 결과를 100 % 정확히 예측할 경우, 모든 점은 대각선상에 위치하게 된다. 즉, 대각선상에 점들이 모여 있을 경우, 예측 정확도가 높다고 판단할 수 있다. 검은색 사각형 심벌은 훈련 데이터에

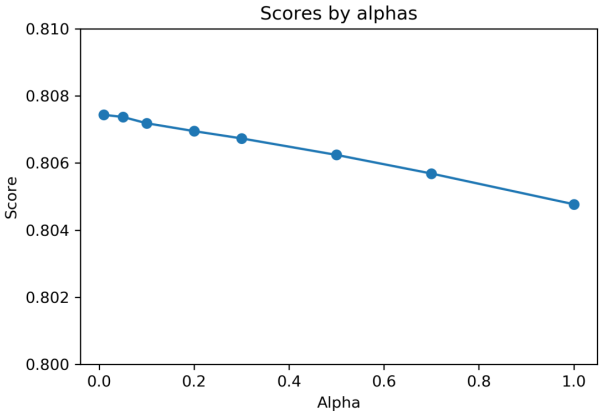


Fig. 4 Hyperparameter tuning in Ridge regression

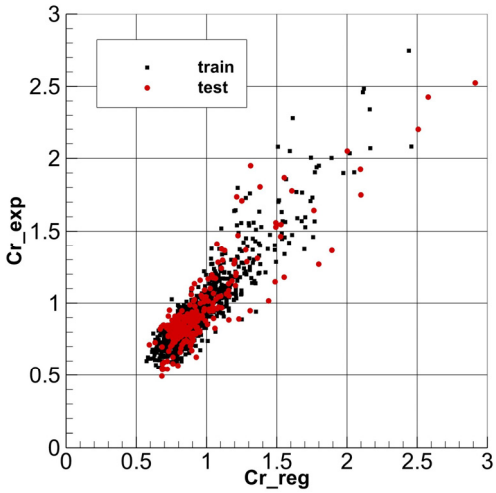


Fig. 5 Cr_reg-Cr plot of Ridge regression

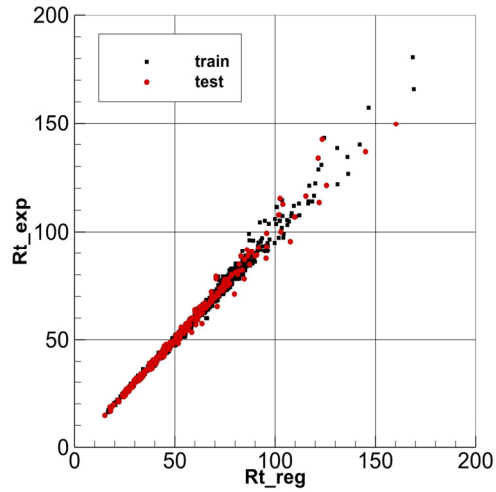


Fig. 6 Rt_reg-Rt plot of Ridge regression

대한 모델 예측 결과를 보이며, 빨간색 원형 심벌은 평가 데이터에 대한 예측 결과를 의미한다. 두 결과 모두 편차에 큰 차이를 보이지 않는 것으로 보아 과대적합이 일어나지 않은 것으로 판단된다. 다만 Fig. 1에서 보이는 바와 같이 고속의 Cr이 큰 영역에서 데이터의 양이 많지 않기 때문에 Cr이 큰 영역에서의 편차는 상대적으로 크게 나타남을 확인할 수 있다. 훈련 데이터에 대한 Cr의 평균 오차는 약 8.4 %로 계산되었고, 평가 데이터에 대한 오차는 약 10.9 %이다. Fig. 6은 예측된 Cr을 모형선의 Rt로 환산한 결과를 보인다. 전저항의 오차는 훈련 데이터가 약 2.0 %, 평가 데이터가 약 2.5 % 수준을 보였다. 최대 오차는 약 13 %로 계산되었다.

3.3 서포트 벡터 머신

서포트 벡터 머신(SVM)은 강력하고 널리 사용되는 학습 알고리즘 중 하나이다. SVM은 데이터의 분류(classification)에 주로 사용되는 방법이지만 유사한 방식으로 회귀 문제(SVR)에도 적용이 가능하다 (Smola & Scholkopf, 2003). SVM은 데이터를 구분할 수 있는 경계면을 찾는 문제로 두 경계면 사이의 거리(margin)를 최대로 하는 경계면을 찾는다(Fig. 7). 복잡한 데이터의 경우, 이러한 선형 경계를 찾는 것이 통상 불가능하기 때문에 일부 데이터는 두 경계면 사이에 존재해도 되도록 제약을 완화하는 소프트 마진(soft margin)을 적용하며, 고차원 맵핑을 통해서 경계면을 찾기 위해서 커널 트릭(Kernel trick)을 사용하기도 한다. 사용되는 커널로는 선형(linear), 다항(polynomial), 방사 기저 함수(radial based function; RBF), 시그모이드(sigmoid) 등이 있으며, 특별히 독립변수의 수가 많은 경우(이러한 경우 선형 커널이 추천됨)를 제외하고, 일반적으로 입력 차원에 구애받지 않고 맵핑이 가능한 방사 기저 함수가 주로 사용되며, 본 연구에서도 방사 기저 함수를 적용하였다.

소프트 마진의 패널티를 부과하는 변수 C, 두 경계면의 거리를 조절하는 epsilon에 대하여 하이퍼파라미터 튜닝을 수행하였으며, Fig. 8에 그 결과를 보인다. RBF의 크기를 조절하는 γ는 1/(독립변수의 개수×독립변수의 분산)으로 설정하는 방식을 사용하였다. Table 3은 테스트에 사용된 파라미터 세트를 보인다. Fig. 8과 같이 epsilon에 대해서는 그 크기가 커질수록 검증

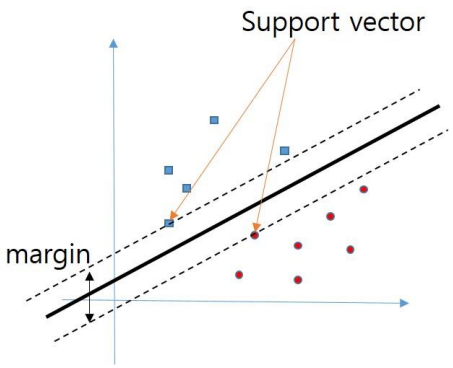


Fig. 7 Support Vector Machine

점수가 작아지는 경향을 보이며, 소프트 마진 패널티는 50 이상에서 수렴하는 결과를 보인다. 본 연구에서는 C를 50, epsilon을 0.01로 설정하고 최종 모델을 구성하였다.

Fig. 9는 Cr에 대한 회귀 추정 결과를 보인다. 훈련 데이터에 대한 모델 적합도는 릿지 회귀에 비하여 상당히 높은 반면, 평가 세트에 대해서는 유사한 정도의 적합도를 보이고 있다. 과대적합이 일어난 것을 간접적으로 확인할 수 있다. 또한 일부 데이터에 대해서는 과도한 과소평가 결과도 보이고 있다. 훈련 데이터에 대한 평균 오차는 약 1.4 %, 평가 데이터에 대한 오차는 약 9.5 %로 계산되었다. 전저항의 오차는 훈련 데이터가 약 0.4 %, 평가 데이터가 약 2.3 % 수준을 보였다. 최대 오차는 약 20.7 %로 계산되었다.

Table 3 Hyperparameter sets for support vector regression

No.	C	epsilon
0	1	0.01
1	1	0.05
2	1	0.1
3	5	0.01
4	5	0.05
5	5	0.1
6	10	0.01
7	10	0.05
8	10	0.1
9	50	0.01
10	50	0.05
11	50	0.1
12	100	0.01
13	100	0.05
14	100	0.1
15	500	0.01
16	500	0.05
17	500	0.1

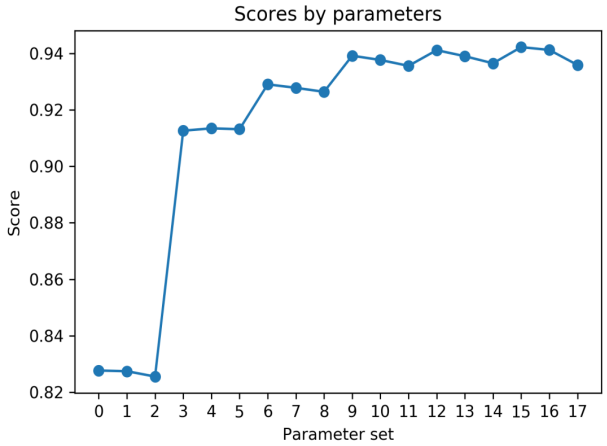


Fig. 8 Hyperparameter tuning in support vector regression

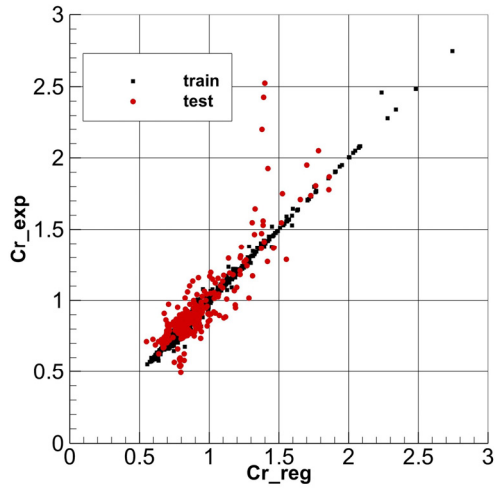


Fig. 9 Cr_reg-Cr plot of support vector regression

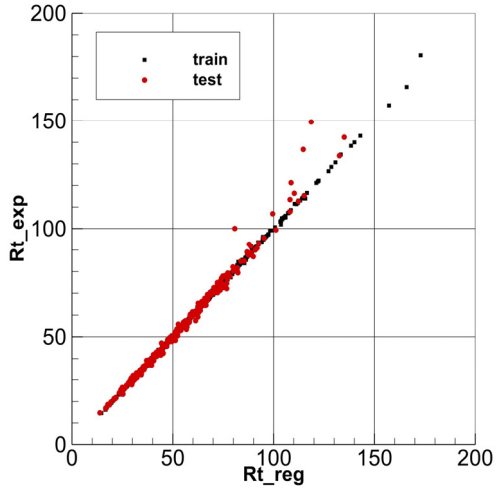


Fig. 10 Rt_reg-Rt plot of support vector regression

3.4 랜덤 포레스트

랜덤 포레스트(random forest; RF)는 결정 트리(decision tree) 기반의 방법으로 입력 데이터 스케일에 영향을 받지 않아 입력 데이터의 정규화나 표준화 같은 전처리 과정이 필요 없는 방법이다. 결정 트리 방법은 하나의 회귀 결과를 얻기 위해서 여러 번의 질문 과정을 통해서 데이터를 분류하고, 최종 잎(leaf node)에 해당하는 데이터의 평균값을 이용하여 결과를 도출한다. Fig. 11은 결정 트리를 이용한 회귀 과정의 개념도를 보인다. 일반적으로 결정 트리는 과대적합이 일어나기 쉬운 방법으로 무작위 트리의 조합(ensemble)을 이용해서 과대적합의 위험도를 낮추는 랜덤 포레스트 방법 (Breiman, 2001)이 주로 사용된다.

랜덤 포레스트의 성능 튜닝은 주로 트리의 개수(n_estimation)로 이루어지며, 본 연구에서는 [10, 15, 20, 25, 30, 100, 200, 300, 400, 500] 총 10개의 트리 개수에 대하여 튜닝을 수행하였다. Fig. 12는 그 결과를 보인다. 트리 개수 400에서 가장 좋은 검증 점수를 보이고 있다. Fig. 13은 최종 모델에서 각 변수들의 중요도를 표시한다. Cb에 crpoly를 곱한 변수의 중요도가 가장 높게 나타났으며, 그 다음으로 선수부의 Cb에 crpoly를 곱한 변수의 중요도가 높게 나타나서, 현재 모델에서는 저속비대선의 Cb와 속도에 관련된 변수들이 중요 인자임을 알 수 있다.

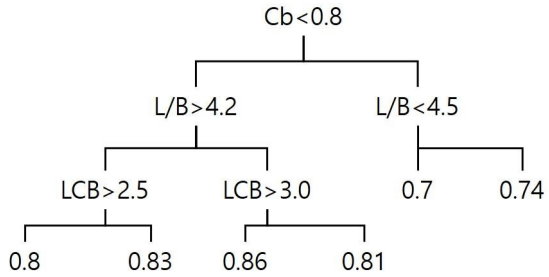


Fig. 11 Decision tree-based regression

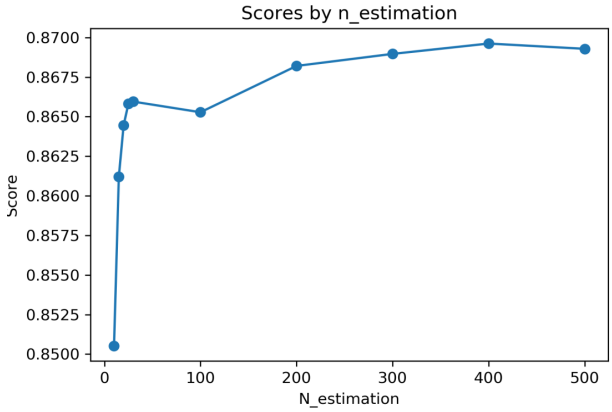


Fig. 12 Hyperparameter tuning in random forest

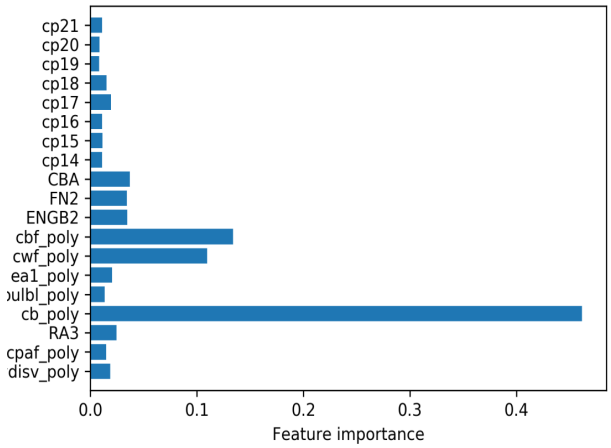


Fig. 13 Feature importance of the RF model

Fig. 14와 15는 각각 랜덤 포레스트에 의한 Cr과 Rt의 예측 결과를 보인다. SVR과 유사한 결과를 보이나, SVR에 비해 과대적합은 줄어든 것으로 보인다. 과도한 예측 케이스도 사라진 것을 확인할 수 있다. 훈련 데이터에 대한 Cr의 평균 오차는 약 1.7 %, 평가 데이터에 대한 오차는 약 9.5 %로 나타났다. 전저항의 오차는 훈련 데이터가 약 0.5 %, 평가 데이터가 약 2.2 % 수준을 보였다. 최대 오차는 약 13.0 %로 계산되었다.

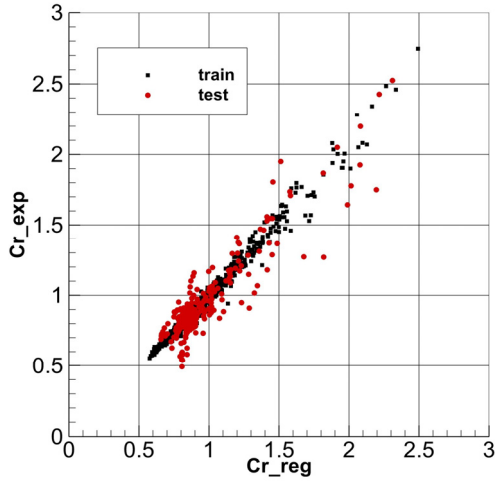


Fig. 14 Cr_reg-Cr plot of random forest

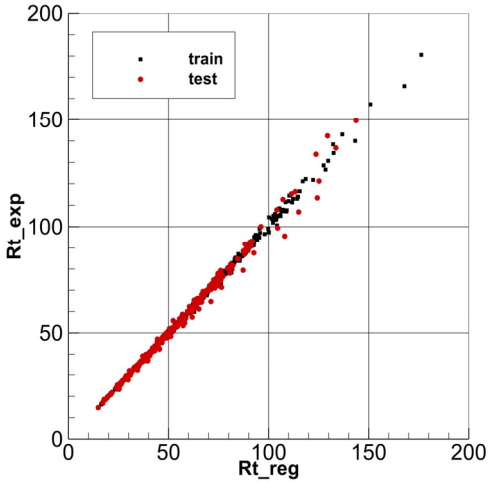


Fig. 15 Rt_reg-Rt plot of random forest

3.5 신경망

신경망(neural network; NN) 모델은 최근 딥러닝과 관련하여 가장 비약적인 발전을 이룬 기계학습 모델의 하나이다. 신경망은 Fig. 16과 같이 입력값(input layer)의 선형 결합으로 이루어진 식을 활성화 함수(activation function)에 적용하여 다음 층의 값을 계산하고, 이와 같은 은닉층(hidden layer)을 여러 겹 구성하여 최종 출력값(output layer)을 계산하는 방식을 사용한다. 그림에서 각각의 화살표에 대응하는 가중치를 설정하고, 최종 출력값과 실제값(정답)과의 차이를 줄이는 방향으로 각각의 가중치들을 조정하는 역전파(back propagation) 방법을 활용하여 모델 최적화를 수행한다. 신경망은 대량의 데이터에 내재된 정보를 잡아내고 복잡한 모델을 만들 수 있다는 장점을 가진 반면, 학습 시간이 길고, 모든 데이터에 대한 표준화를 수행하여야 잘 작동하는 특징을 갖는다. 또한 은닉층의 수, 각 층의 유닛수 등 모델 구성에 대한 다양한 조합이 존재하며, 좋은 모델을 찾기 위해서는 수많은 테스트를 거쳐야 하는 단점도 존재한다.

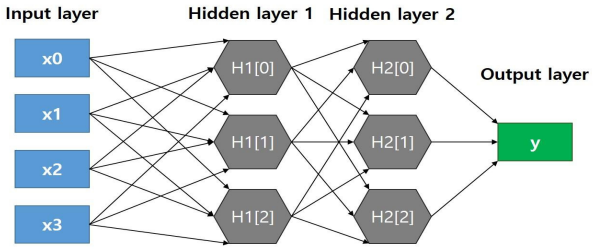


Fig. 16 Neural network with multiple hidden layers

본 연구에서는 신경망 모델에 대한 광범위한 튜닝은 실행하지 않고, 활성화 함수로 속도가 빠르고 널리 사용되는 ReLU 함수를 채용하였고, 가중치 갱신을 위한 최적화 방법으로 adam 기법 (Kingma & Ba, 2015)을 적용한 후, 최적 epoch수(전체 데이터 세트를 모두 사용하여 학습한 횟수)를 찾는 방식을 이용하였다. 은닉층의 수와 유닛수는 깊이가 깊고 얕은 모델, 유닛수가 많고 적은 모델로 분류하여 테스트를 진행하였으며, 학습은 50 epoch 동안 오차가 개선되지 않으면 도중에 학습을 중지하는 방식을 사용하였다. Fig. 17에서와 같이 은닉층의 깊이에는 결과가 크게 변하지 않으며, 각 은닉층의 유닛수가 많은 쪽에서 훈련 데이터에 대한 오차(MSE; Mean Squared Error)가 더 작게 나타났다. 은닉층의 개수가 2이고, 유닛수 128인 경우를 최종 모델 구성에 사용하였다.

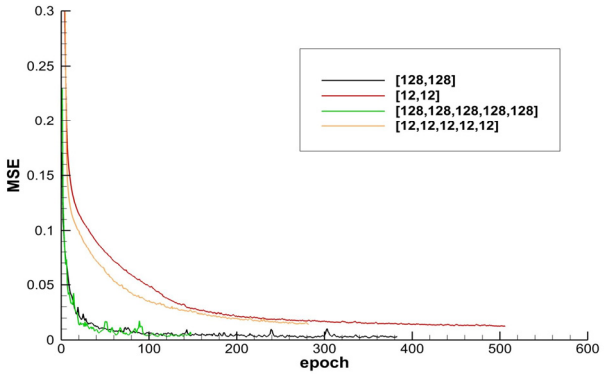


Fig. 17 Training results according to layer construction

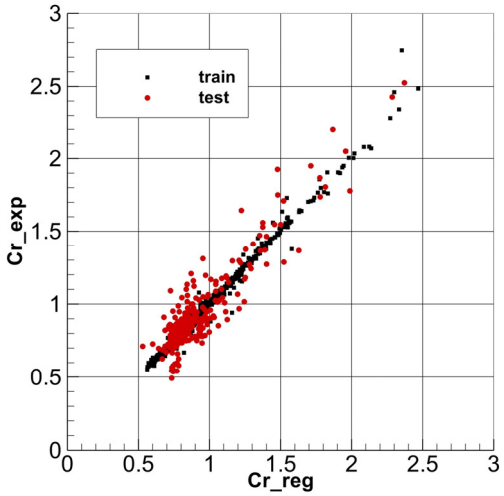


Fig. 18 Cr_reg-Cr plot of neural network

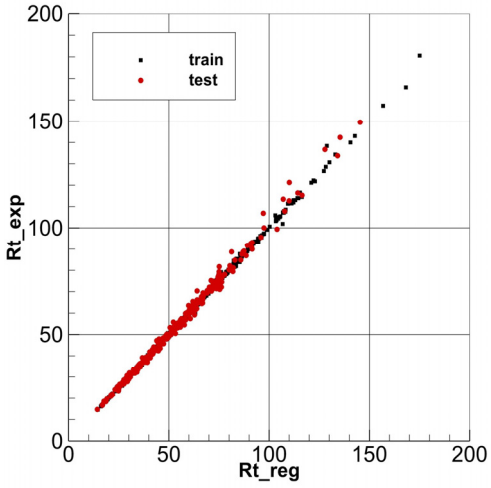


Fig. 19 Rt_reg-Rt plot of neural network

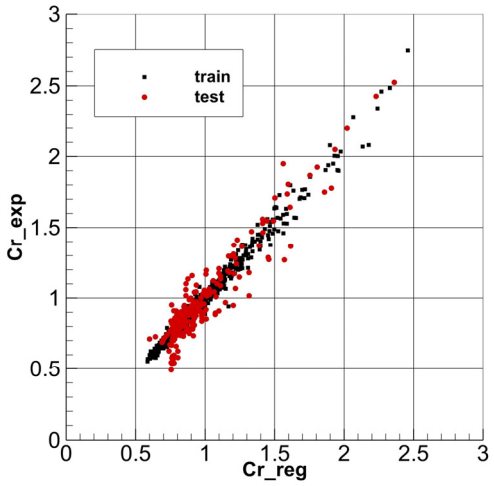


Fig. 20 Cr_reg-Cr plot of ensemble model

Fig. 18과 19는 신경망을 이용하여 예측한 Cr과 Rt를 보인다. 훈련 데이터에 대한 Cr의 평균 오차는 약 1.4 %로 랜덤 포레스트 대비 약간 향상되었으며, 평가 데이터에 대한 오차는 약 9.7 %로 랜덤 포레스트보다 미세하게 과대적합된 것으로 판단된다. 전저항의 오차는 훈련 데이터가 약 0.4 %, 평가 데이터가 약 2.3 % 수준을 보였고 최대 오차는 약 9.3 %로 계산되었다.

3.6 앙상블 학습

앙상블 방법은 여러 기계학습 모델을 하나의 메타 모델로 연결하여 개별 모델보다 더 좋은 일반화 성능을 내기 위하여 사용된다 (Raschka & Mirjalili, 2019). 일반적으로 회귀 모델에서는 각 모델의 예측치를 가중치를 적용하여 평균하는 방식을 사용한다. 본 연구에서는 앞서 소개한 4개의 방법(릿지 회귀, 서포트 벡터 머신, 랜덤 포레스트, 신경망)을 통해 예측된 Cr을 가중 평균하여 최종 Cr을 계산하였다. 각 모델의 가중치는 테스트 점수가 높고, 과대평가된 예측치를 보이지 않은 랜덤 포레스트와 신경망에 더 높게 설정하여 [0.2, 0.1, 0.35, 0.35]로 결정하였다. Fig. 20과 21은 신경망을 이용하여 예측한 Cr과 Rt를 보인다. 훈련 데이터에 대한 Cr 평균 오차는 2.4 %로 랜덤 포레스트나 신경망 개별 모델보다 좋지 않은 결과를 보이고 있으나, 평가 데이터에 대한 오차는 8.4 %로 향상되는 것을 확인할 수 있다. 이는 개별 모델 대비 과대적합이 줄어든 것으로 해석할 수 있다. 특히 높은 Fn에서의 예측 정확성이 향상된 것으로 판단된다. 전저항의 오차는 훈련 데이터가 약 0.6 %, 평가 데이터가 약 1.9 % 수준을 보였고 최대 오차는 약 8.1 %로 모든 모델 중 가장 좋은 결과를 보였다.

3.7 모델 검증

가장 좋은 결과를 보인 앙상블 모델을 이용하여 선형 변환에 따른 잉여저항계수 예측 성능에 대한 검증을 수행하였다. Kim et al. (2019)은 선형 회귀식의 검증을 위하여 Supramax 60K

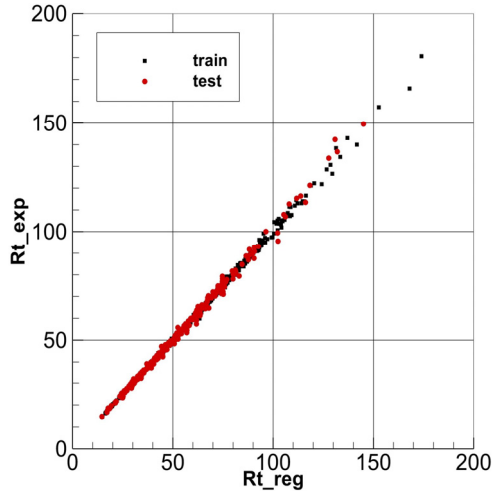


Fig. 21 Rt_reg-Rt plot of ensemble model

선형에 대한 파생 선형들을 설계하고 이에 대한 잉여저항계수를 추정하여 CFD 결과 및 모형시험 결과와 비교한 바 있다. 본 연구에서는 이 선형들에 대한 잉여저항계수를 추정하고 CFD 결과와 비교를 수행하였다. 검증에 사용된 선형은 Supramax 기준 선형과 LCB 위치를 변경한 M1 선형, 유사 실적선의 선형을 Supramax 선형 제원에 맞춘 M2 선형, M2 선형의 LCB 위치를 조정한 M3 선형, M3 선형의 선미 부피를 조정한 M4 선형, 이렇게 총 5척이다 (Kim et. al, 2019).

Fig. 22는 선박해양플랜트연구소(KRISO)의 in-house CFD 코드인 WAVIS를 사용하여 각 선형에 대한 잉여저항계수를 계산한 결과이며, Fig. 23은 Kim et al. (2019)의 선형 회귀식에 의한 결과, Fig. 24는 앙상블 모델을 사용하여 선형변수만으로 예측한 잉여저항계수값을 보인다. 앙상블 모델이 CFD 결과와 정략적으로 일치하지는 않지만, 각 선형에 따른 경향은 어느 정도 반영하고 있음을 확인할 수 있다. 특히 회귀 모델들이 비교적 데이터가 부족한 높은 Fn 영역에서 CFD 결과와 차이를 보이는 것을 확인할 수 있다. 선형 회귀 (Kim et al., 2019)에 의한 예측이 CFD 결과와 더 유사한 경향을 주고 있으나, 이 모델

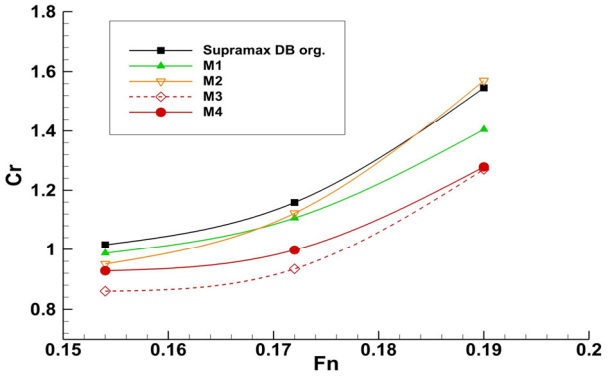


Fig. 22 Cr predictions based on CFD (Kim et al., 2019)

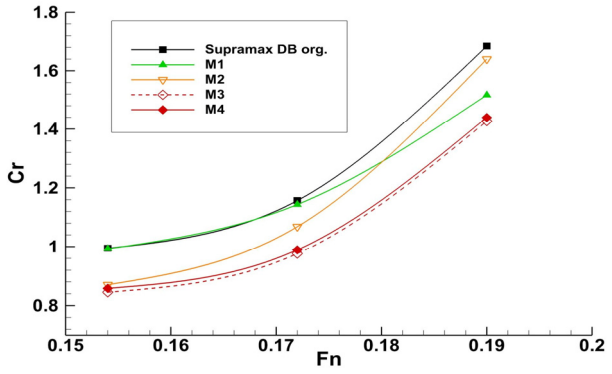


Fig. 23 Cr predictions based on the linear regression (Kim et al., 2019)

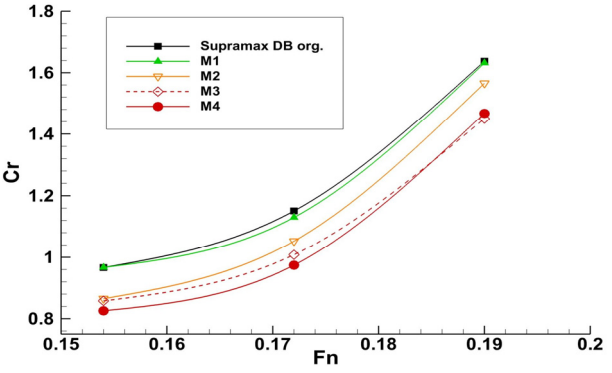


Fig. 24 Cr prediction based on the present ensemble model

은 학습시 모든 데이터를 사용하여 학습이 진행되었고, 회귀식 도출시 데이터 분류를 Cb와 속도 영역에 따라 세분화하여 도출한 점, 사용된 독립변수에서 현재 모델과 차이점이 있다. 현재 모델은 분석시 복잡한 데이터 분류를 사용하지 않고 도출한 모델인 점을 감안하면, 활용 가능성이 있다고 판단된다.

4. 결 론

본 연구에서는 KRISO의 모형시험 데이터베이스 중 벌크선과 탱커선을 포함하는 저속비대선 선형에 대하여 기계학습 기법을 활용하여 선형 변수로부터 잉여저항계수를 예측하는 추정 모델을 도출하고 검증을 수행하였다. 기계학습 모델로서 릿지 회귀, 서포트 벡터 머신, 랜덤 포레스트, 신경망에 대하여 검토를 수행하였다. Table 4는 각 모델에 대한 훈련 데이터와 평가 데이터에 대한 Cr 추정 오차를 나타낸다. 릿지 회귀의 경우, 단순한 모델 특성상 과소적합의 경향이 다소 보였고, 다른 모델들은 과대적합의 특성이 나타난 것을 확인할 수 있다. SVR의 경우는 일부 평가 데이터에 대해서 과도한 예측 결과를 주는 것이 관찰되었다. 최종 모델로는 훈련 데이터에 대해서는 개별 모델보다 예측 점수가 떨어지지만, 평가 데이터에 대해서 더 좋은 결과를 주어, 과대적합이 억제된 앙상블 모델을 선정하고, Supramax 선형과 이를 변형한 파생 선형에 대하여 Cr 예측을 수행하였다. CFD 결과, 그리고 기존의 선형 회귀 결과와의 비교를 통해서 기존 선형 회귀 분석보다 단순한 데이터 처리 방식으로도 유사한 결과를 도출할 수 있음을 확인하였다.

Table 4 Summary of Cr prediction error for each machine learning model

Cr error(%)	Ridge	SVR	RF	NN	Ensemble
training data	8.4	1.4	1.7	1.4	2.4
test data	10.9	9.5	9.5	9.7	8.4

본 연구에서는 초기 선형 설계시 사용할 수 있는 몇 개의 선형 변수를 이용하여, 각 속도별 잉여저항계수를 통계에 기반하여 예측할 수 있는 기계학습 모델에 대한 검토를 수행하였다. 릿지 회귀를 제외한 복잡한 비선형 모델들은 과대적합의 특성을 갖고 있음을 확인할 수 있었으며, 이런 모델들의 사용에 주의가 필요할 것으로 판단된다. 일부 선형 변수들만 사용해서 잉여저항계수의 정확한 예측에는 한계가 따르며, 통계 분석 특성상 모집단의 범위를 벗어난 선형에 대한 예측 결과는 신뢰할 수 없음에 주의가 필요하다. 기계학습 모델은 양질의 데이터가 뒷바침되어야 그 성능이 보장될 수 있기 때문에 충실한 선형 및 모형시험 결과, CFD 결과에 대한 데이터베이스화가 지속되어야 하며, 또한 본 연구에 사용된 선형 변수 외에 다양한 변수에 대한 검토도 필요할 것으로 생각된다. 정량적으로 더 근접한 예측 결과를 위해서는 선형 오프셋 데이터의 학습에 의한 기계학습 모델 도출 필요성도 있다고 판단된다.

후 기

본 논문은 선박해양플랜트연구소 주요사업 “기계학습을 이용한 선박의 소요 마력 성능 추정 기반 기술 연구” 및 “첨단운송체의 항내 운항성능향상을 위한 축척효과를 고려한 운항제어원천기술 개발”로 수행된 결과입니다 (PES3710, PES3410).

References

Breiman, L., 2001. Random forests, *Machine Learning*, 45, pp.5-32.

Cho, Y.I. et al., 2019. Resistance estimation of a ship in the initial hull design using deep learning. *Korean Journal of Computational Design and Engineering*, 24(2), pp.203-210.

Gertler, M., 1954. *A reanalysis of the original test data for the Taylor standard series*. Navy Department The David W. Taylor Model Basin, Report 806.

Guldhammer, H.E. & Harvald, Sv. Aa., 1965. *Ship resistance effect of form and principal dimensions*, Akademisk Forlag, Copenhagen.

Holtrop, J., 1984. A statistical re-analysis of resistance and propulsion data. *International Shipbuilding Progress*, 31, pp.272-276.

Holtrop, J. & Mennen, GGJ., 1978. A statistical power prediction method. *International Shipbuilding Progress*, 25, pp.253.

Holtrop, J. & Mennen, GGJ., 1982. An approximate power prediction method. *International Shipbuilding Progress*, 29, pp.166-170.

Kim, H.C., & Park, H.G., 2015. Practical application of neural networks for prediction of ship's performance factors. *Journal of Ocean Engineering and Technology*, 29(2), pp.111-119.

Kim, H.J., Chun, H.H., & Choi, H.J., 2007. Development of CFD based stern form optimization method. *Journal of the Society of Naval Architects of Korea,* 44(6), pp.564-571.

Kim, J. et al., 2011. Development of a numerical method for the evaluation of ship resistance and self-propulsion performances. *Journal of the Society of Naval Architects of Korea,* 48(2), pp.147-157.

Kim, Y.C. et al., 2019. Prediction of residual resistance coefficient of low-speed full ships using hull form variables and model test results. *Journal of the Society of Naval Architects of Korea*, 56(5), pp.448-457.

Kingma, D.P., & Ba, L.J., 2015. Adam: A method for stochastic optimization, *International Conference on Learning Representations (ICLR)*, San Diego, USA, 7-9 May 2015.

Lap, A.J.W., 1956. Resistance (Fundamentals of ship resistance and propulsion). *International Shipbuilding Progress*, 3(24), pp.441.

Park, J.H., Choi, J.E., & Chun, H.H., 2015. Hull-form optimization of KSUEZMAX to enhance resistance performance. *International Journal of Naval Architecture and Ocean Engineering*, 7(1), pp.100-114.

Park, S.H., Lee, S.B., & Lee, Y.M., 2014. Study on the estimation of the optimum trims in container carriers by using CFD analysis of ship resistance. *Journal of the Society of Naval Architects of Korea,* 51(5), pp.429-434.

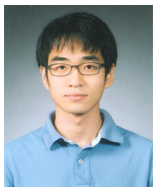
Raschka, S., & Mirjalili, V., 2019. Python machine learning 2nd ed., Packt.

Rifkin, R.M., & Lippert, R.A., 2007. Notes on regularized least squares. *Computer Science and Artificial Intelligence Laboratory Technical Report*.

Smola A.J., & Scholkopf, B., 2003. A tutorial on support vector regression, Statistics and Computing, 14, pp.199-222.

Taylor, DW., 1933. *Speed and power of ships*. Washington, DC: Press of Ransdell.


김 유 철


양 경 규


김 명 수


이 영 연


김 광 수