J. Comput. Chem. Jpn., Vol. 22, No. 2, pp. 31–33 (2023)

速報

分子軌道エネルギーを用いた機械学習によるエントロピーの予測

結城 敬史 , 中原 和加奈 , 寺前 裕之 *

城西大学理学部 〒350-0295，埼玉県坂戸市けやき台1-1
teramae@gmail.com

(Received: August 30, 2023; Accepted for publication: October 25, 2023; Online publication: January 24, 2024)

The values of the entropy of 148 small organic molecules have been estimated by machine learning with only molecular orbital energies as the explanatory variables. Out of 148 molecules，we used 104 molecules for the training set and 44 molecules for the test set. We used 139 regression methods of R/caret package for machine learning. We evaluated values by RMSE (Root Mean Squared Error) and R^2 (coefficient of determination). From those evaluation, xgbLinear (eXtreme Gradient Boosting) and RRFglobal (Regularized Random Forest) are considered better than other regression methods. It has been proved that the entropy can be predicted by the molecular orbital energies only.

Keywords : entropy，molecular orbital energy，machine learning，prediction of entropy，explanatory variables

1　はじめに

情報社会において計算科学は様々な分野に応用されつつあり，今後も発展が期待される. 中でも機械学習は与えられたデータから回帰，分類を行うことにより集団の性質や予測を立てることを可能とするため，他の分野と同様に化学分野においても機械学習によってデータ予測が期待できる. 物性を予測するにあたって化合物の種類は数多くあり，また検討すべき要素は多く存在するため，より一層複雑となる. そのため様々な物性を少ない要素で説明できることが望ましい. 近年，我々は物性の予測に関し，分子軌道エネルギーを説明変数として機械学習を行うことでより単純な形で予測が行えるのではないかと考え研究を行ってきた [1, 2].

我々は以前にR言語のcaretパッケージ [3]を使用しrandom forest (rf)，generalized linear model (glm)，least angle regression (lars)，projection pursuit regression (ppr)の4種類の回帰法を用いて説明変数としての分子軌道エネルギーの数を2，4，6，8，10，12，14，16，18，20として全150分子のエントロピー値を目的変数とした機械学習を行い，トレーニングデータとテストデータを含む相関係数Rにより評価を行った結果，強い相関関係を持つことを示した [4].

本研究ではさらに機械学習に用いる回帰法を136種類に増やし 説明変数としての軌道数は2，4，6，8，10とし，二乗平均平方根誤差誤差(RMSE値)と，決定係数(R^2)を指標として，予測にするに適した回帰法と説明変数としての軌道数の探索を行ったので報告する.

2　計算方法

アルカン，芳香族などを含む有機分子，148種について気体のエントロピーのデータを化学便覧 [5]より用意し，対応する分子の分子構造をGaussian16プログラム[6]によりHF/6-31G(d，p)レベルで最適化を行い，振動数計算を行うことでエネルギー極小値にあることを確かめた後に，分子軌道エネルギーのデータを作成した. 説明変数として軌道数は2，4，6，8，10の5種類を使用した. 作成したデータをトレーニングデータとテストデータに分け，トレーニングデータについて5-fold cross-validationを行った. その後，それらを用い，各々の回帰法で計算させた. なお，ハイパーパラメータはcaretのデフォルトのまま使用した.

3　結果

エラーの生じなかった99種類の回帰法から，RMSE値，R^2値を用いて絞り込みを行った. RMSE値が0.06以上の結果を除外し，決定係数R^2が値0.8未満を除外した結果11種類の回帰法が選択された. 選択した回帰法は以下の11種類である. qrf (Quantile Random Forest)，RRF (Regularized Random Forest)，Rborist (Random Forest)，parRF (Parallel Random Forest)，ranger (Random Forest)，rf (Random Forest)，xgbLinear (eXtreme Gradient Boosting)，xgbDART (eXtreme

DOI: 10.2477/jccj.2023-0025

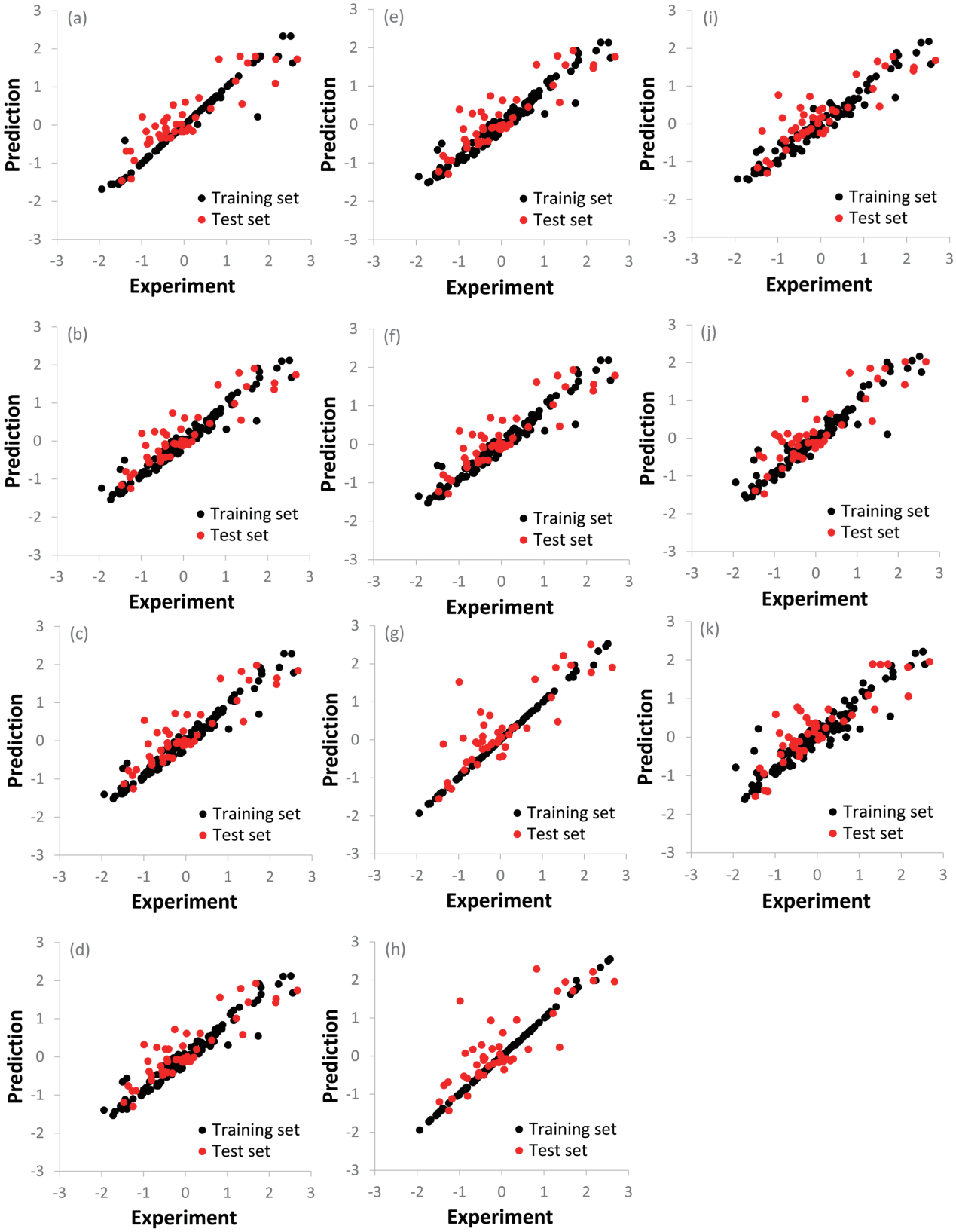

Figure 1. Correlation Plots of values of entropy of experiment and prediction with respect to training and test set for 11 regression methods. Regression methods are (a) qrf、(b) RRF、(c) Rborist、(d) parRF、(e) ranger、(f) rf、(g) xgbLinear、(h) xgbDART、(i) RRFglobal、(j) bstTree、and (k) kknn、respectively.

Table 1. Coefficients of determination (R^2), root mean squared errors (RMSE) and numbers of molecular orbitals for 11 different regression methods.

method	R^2 (train+test)	RMSE (test)	Number of molecular orbital
qrf	0.900	0.0538	10
RRF	0.897	0.0538	10
Rborist	0.895	0.0540	10
parRF	0.895	0.0541	10
ranger	0.893	0.0537	10
rf	0.892	0.0540	10
xgbLinear	0.891	0.0579	8
xgbDART	0.888	0.0590	10
RRFglobal	0.875	0.0589	8
bstTree	0.867	0.0536	10
kknn	0.841	0.0555	10

Gradient Boosting), RRFglobal (Regularized Random Forest), bstTree (Boosted Tree), kknn (k-Nearest Neighbors),

またそれぞれのトレーニング結果を表にしたものを示す. Table 1で使用したR^2はトレーニングデータとテストデータの全体の値，RMSEはテストデータについてのものである. またそれぞれ11種類の回帰法による機械学習結果をFigure 1に示す.

選択した11種類のうちxgbLinearとRRFglobalの2種類が軌道数 8 でRMSEが極小になったがそれらを除く 9種類, qrf, RRF, Rborist, parRF, ranger, rf, xgbDART, bstTree, knnで軌道数10となり極小値は得られなかった.

4 考察

上述のように今回の結果では，R^2値，RMSE値が説明変数の数8で極小となったxgbLinearとRRFglobalが分子軌道エネルギーを用いた機械学習によるエントロピーの予測において優れていると考えられる. それ以外の回帰法に関して，説明変数の数を10より増やして機械学習すると最適となるのか，あるいは説明変数の数を増やしたとしても単に過学習が起こるに留まるかが疑問点として挙げられる. 説明変数の数を増やすことで，反応に最も関与すると考えられるHOMO(最高被占軌道) LUMO(最低空軌道)からより遠い軌道の分を多く考慮しなくてはならないというのは物理的観点から考えると疑問であるが，説明変数の数10以上に関しては今後の検討課題であろう.

選択された上位11種類の回帰法のうち，7種類がrandom forest系であったのは興味深い. 一方でNeural network系のものは1種類も含まれていなかった. Neural network系のものはもう少しトレーニングセットの数が必要なのかもしれず，今後の検討課題であろう.

またそれぞれの機械学習の結果のうち集団から最も離れた分子を特定した. bstTreeの，dinitromethaneを除く10の回帰法でtrifluoromethaneが最も離れていた. Trifluoromethaneはフッ素を多く含み，ハロゲンを多く含むものは構造に影響を与えるため，構造の差が集団から離れた要因として関係しているのではないかと考えている.

より多くの回帰法を試すため，各々の回帰法のハイパーパラメータやcross validationのコマンドについて，細かく設定はしなかった. 特にcross validation については分割，評価ともにcaretの機能に任せた. 今後は精度の向上や可視化のため，パラメータチューニングや交差検証の評価値，標準偏差の算出など前向きに検討したい.

5 結論

本研究では，分子軌道エネルギーを説明変数として用いて148種類の有機分子のエントロピー値に関する機械学習を行った. 回帰法136種類のうちrandom forest系がテストデータも含む全体のR^2値やテストデータに対するRMSE値に対して良好な結果を与えた. ただし，軌道エネルギー数10以下では, RRFglobalとxgbLinear以外はこれらの値の極小値を与えることは出来なかった．軌道エネルギーの物理的意味を考えると全般的にこのRRFglobalとxgbLinearの2種類が優れていると思われる.

参考文献

[1] H. Teramae, M. Xuan, J. Takayama, M. Okazaki, T. Sakamoto, *AIP Conf. Proc.*, **2611**, 020007 (2022). DOI:10.1063/5.0119589

[2] H. Teramae, X. Meiyan, J. Takayama, M. Okazaki, T. Sakamoto, *J. Comp. Chem*, *Jpn.*, **21**, 103-105 (2022)

[3] https://cran.r-project.org/web/packages/caret/caret.pdf

[4] H. Teramae, unpublished results.

[5] The Chemical Society of Japan, Handbook of Chemistry: Pure Chemistry, 5th ed., Maruzen, 2004.

[6] Gaussian 16, Revision B.01, M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox, Gaussian, Inc., Wallingford CT, 2016.