6S-02

木構造に基づく機械学習モデルによるホスフィンの電子的性質予測

◯高橋青玄 † 山口拓也 † 柳村海希 † 山中克久 † 吉田尚恵 † 是永敏伸 †
岩手大学 †

1 序論

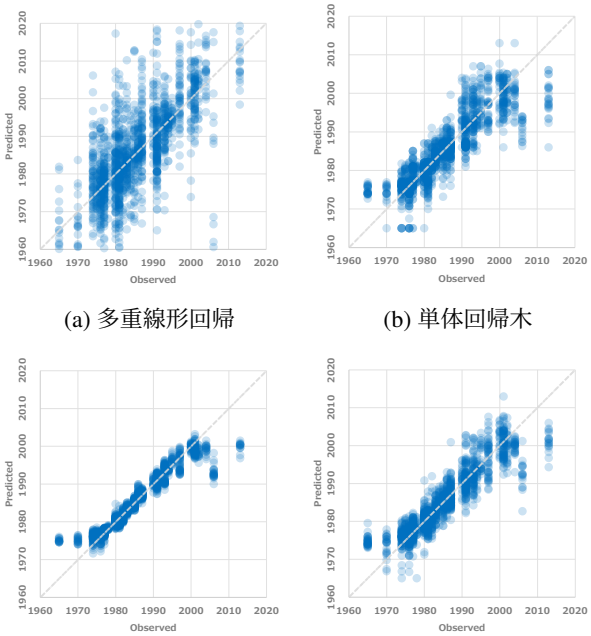
ホスフィン配位子を持つ金属錯体触媒は、薬品の開発や材料の製造等、化学産業で広く用いられている。ホスフィン配位子を有する金属錯体触媒の電子的性質は、その触媒の性能に密に関係し、その電子的性質は、カルボニル伸縮振動と深く関連する。よって、カルボニル伸縮振動を知ることは、その触媒の性能を評価する上で非常に重要である。しかしながら、カルボニル伸縮振動を測定するには実際に触媒を合成する必要がある。一方、近年、触媒開発をはじめとした化学分野において機械学習が注目を集めている。カルボニル伸縮振動の予測を行う機械学習モデルを構築することができれば、試行錯誤的な合成を減らし、効率的な触媒開発に寄与することができる。化学分野において機械学習を利用した研究の一例として Ahneman ら [1] の報告が挙げられる。Ahneman ら [1] は、バックワルドハートウィッグアミノ化における収率予測に対して、ランダムフォレストが高い予測を行うことを示した。他にも、Takigawa ら [2] は、金属触媒における d バンド中心予測において勾配ブースティング決定木の有用性を示した。本研究では、木構造に基づく機械学習モデルが、カルボニル伸縮振動予測においてどの程度の性能を有するのかを調査する。

2 使用した機械学習モデルとデータセット

本研究では、カルボニル伸縮振動予測を行う機械学習モデルとして、多重線形回帰 (Linear)、回帰木 (TREE)、ランダムフォレスト (RF)、勾配ブースティング決定木 (GBDT) の 4 つを使用する。本研究では、主に木構造に基づく機械学習モデルを採用しているが、比較対象として線形なモデルである多重線形回帰を採用している。回帰木、ランダムフォレスト、勾配ブースティング決定木の 3 つのモデルにおけるハイパーパラメータは、適当な範囲において 5 分割交差検証グリッドサーチを行い、最適なパラメータを選択した。検証した値の範囲を Table 1 に示す。

モデル	ハイパーパラメータ
単体回帰木 (TREE)	max_depth $\in [3,4,5,6]$
ランダムフォレスト (RF)	max_depth $\in [3,4,5,6]$, forest_size$=100$ select_feature_num $\in [10,20,...,70]$
勾配ブースティング決定木 (GBDT)	max_depth $\in [3,4,5,6]$ tree_num $\in [50,100,200,350,500]$ learn_rate $\in [0.01,0.05,0.1,0.5,1.0]$

Table 1: 使用モデルとハイパーパラメータ



(a) 多重線形回帰 (b) 単体回帰木

(c) ランダムフォレスト (d) 勾配ブースティング決定木

Figure 1: モデル別の実測値/予測値プロット

次に、本研究で使用したデータセットについて説明する。本研究では、62 種のホスフィン配位子をデータセットとして使用する。それぞれのホスフィン配位子に対して、DFT 計算によって得られた立体的パラメータ、電子的パラメータ、エネルギー、計 76 種を特徴量とする。配位子を実際に合成し、赤外分光法により測定したカルボニル伸縮振動の実測値を目的変数とする。

3 モンテカルロ交差検証

カルボニル伸縮振動予測における機械学習モデルの評価を行うために本研究ではモンテカルロ交差検証を採用する。モンテカルロ交差検証では以下のように単発のホールドアウト検証を複数回行う。m 個のデータからなるデータ集合を D とする。はじめに、D を、サイズ n のテストデータセット D_{test} とサイズ $m-n$ の訓練データセット D_{train} に分割する。次に、D_{train} を用いて機械学習モデル M を生成する。続いて、D_{test} 内の特徴量を用いて、M によって、D_{test} のカルボニル伸縮振動を予測し、予測値と実測値に対して決定係数 (R^2) と二乗平均平方根誤差 ($RMSE$) を算出する。これを指定した回数だけ繰り返すことで R^2 と $RMSE$ のペアを複数個算出し、モデルの評価を行う。

本研究では、62 個のデータセットを、ランダムにテスト/トレーニングへの分割することを 100 回繰り返し、100 回分の R^2, $RMSE$ のそれぞれの平均をモデルの予測精度とする。

Electronic Properties Prediction of Phosphines using Tree-based Machine Learning Models
†Takahashi Seigen, Takuya Yamaguchi, Miki Yanagiura, Katsuhisa Yamanaka, Hisae Yoshida, and Toshinobu Korenaga (Iwate University)

Table 2: モデル別の評価値

		Linear	TREE	RF	GBDT
R^2	平均	-1.2118	0.7366	0.9071	0.8268
	標準偏差	12.9540	0.0170	0.0599	0.0853
RMSE	平均	10.6317	4.4592	2.5803	3.7086
	標準偏差	56.3538	1.5223	1.1713	1.1404

Table 3: 特徴量の選択をした場合の評価値

		76 種	50 種	20 種	5 種
R^2	平均	0.9071	0.9087	0.8995	0.9139
	標準偏差	0.0599	0.0037	0.0041	0.0042
RMSE	平均	2.5803	2.5318	2.7161	2.4151
	標準偏差	1.1713	1.3809	1.2123	1.3568

4 機械学習モデル別の評価

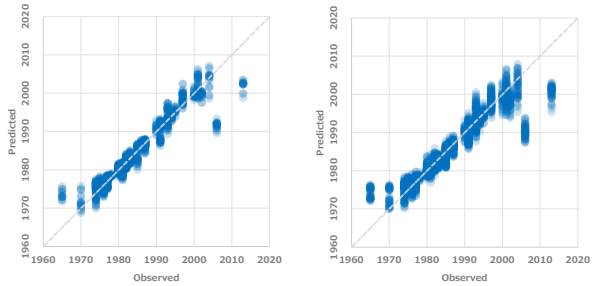
はじめに、各機械学習モデルの評価を行うため、サイズ 62 のホスフィン配位子データセットに対して、テストセット 25%、トレーニングセット 75%の割合でモンテカルロ交差検証を行った。特徴量は 76 種全てを使用した。Figure 1 に 100 回分のモンテカルロ交差検証の試行を全てプロットしたグラフを示す。X 軸は実際に計測された値、Y 軸がモデルにより予測された値である。$X = Y$ となる直線からのずれが予測の誤差である。全試行の評価値の平均と標準偏差を Table 2 に示す。平均 R^2、平均 *RMSE* ともに、ランダムフォレスト、勾配ブースティング決定木の評価値が他 2 つに比べ高く、特にランダムフォレストは平均 $R^2 = 0.9071$ と、非常に高い精度を示した。これよりカルボニル伸縮振動の予測には、ランダムフォレストが適していると結論づけた。

5 特徴量数別評価

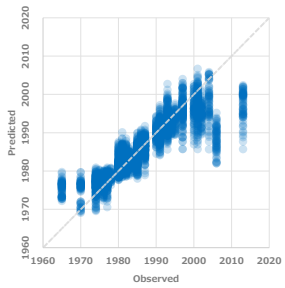
次に、特徴量重要度に基づき、モデルの生成に使用する特徴量の数を調整して実験を行った。特徴量間に高い相関がある場合、特徴量を減らしたとしても、同程度以上の予測を行える可能性がある。ランダムフォレストモデルにおけるジニ重要度をもとに、モデル生成に使用する特徴量の数を調整し、テストデータセット 25%、訓練データセット 75%のモンテカルロ交差検証を行った。機械学習モデルとしてランダムフォレストを使用し、特徴量全 76 種を使用した検証の他に、上位 50 種、20 種、5 種のみを使用した検証を行った。算出された評価値を Table 3 に示す。特徴量を 5 種まで減らしても平均 R^2 は 0.914 と高い精度を示した。

6 訓練データセットのサイズ別評価

最後に、訓練データセットとテストデータセットの比率を 75%/25%、50%/50%、25%/75%と変化させて、上位 5 種の特徴量を用いたランダムフォレストでモンテカルロ交差検証を行った。得られた評価値を Table 4 に示す。訓練データセットの割合を全体の 25% まで絞っても平均 $R^2 = 0.789$ と中程度の精度を示している。



(a) 訓練データ 75%(サイズ 46) (b) 訓練データ 50%(サイズ 31)



(c) 訓練データ 25%(サイズ 15)

Figure 2: 訓練データセットのサイズ別の実測値/予測値プロット

Table 4: 訓練データセットのサイズ別評価値

		訓練データ 75%	50%	25%
R^2	平均	0.9139	0.8771	0.7892
	標準偏差	0.0042	0.0027	0.0112
RMSE	平均	2.4151	3.2115	4.2094
	標準偏差	1.3568	0.8097	1.2097

7 結論

本研究ではホスフィン配位子を持つ金属触媒におけるカルボニル伸縮振動予測を通して、木構造に基づく機械学習モデルの評価を行った。ランダムフォレストが最も高い予測精度を示し、5 種のホスフィン配位子パラメータがあれば十分な予測ができることがわかった。これにより、金属触媒開発における試行錯誤的な合成を減らし、効率的な触媒開発が可能になること期待できる。

References

[1] D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher, and A. G. Doyle. Predicting reaction performance in C–N cross-coupling using machine learning. *Science*, 360(6385):186–190, April 2018.

[2] I. Takigawa, K.-i. Shimizu, K. Tsuda, and S. Takakusagi. Machine-learning prediction of the d-band center for metals and bimetals. *RSC advances*, 6(58):52587–52595, 2016.