有機合成に最適な酵素候補提示のための特徴量エンジニアリングによるEC番号予測

2255018　武藤　克弥

1. はじめに

有機合成化学において, 効率性や環境面から化学反応の設計に酵素を生体触媒として利用する機会が増加している. 酵素は性質によって 4 桁の酵素番号 (Enzyme Commission numbers: EC 番号) で分類されており, 代謝経路の解析における未知酵素の機能特定や, 酵素が触媒する化学反応の探索に EC 番号が用いられてきた. 特に解析や探索における実験コストの削減が期待できることから, 機械学習を用いて EC 番号の予測を行う試みがなされてきた.

本研究では, 従来の物理・化学的特性値と構造的特徴を用いた予測手法を組み合わせ, 化学反応に最適な酵素候補を EC 番号として予測する手法を開発する. それにより, 化学変化の特徴をより詳細に捉え, 従来手法よりも予測精度を向上させることを目指す.

2. EC 番号と有機合成

酵素を分類する EC 番号に対して, 酵素が触媒する化学反応式が 1 つまたは複数登録されており, 酵素を用いた化学反応 (酵素反応) に対して EC 番号がラベル付けされたデータがデータベース上で扱われている. 有機合成では基質から目的の化合物を効率よく得るため, 実験によって酵素を選択するが, 酵素反応から EC 番号を予測することができれば, 合成に用いるべき酵素候補を同じ EC 番号内の酵素製品に絞り込むことができる. そして, 有機合成をより効率的に行うことにつながる.

3. 機械学習による EC 番号予測

機械学習による EC 番号予測では, データベース上にある EC 番号が既に分かっているタンパク質配列や酵素反応データを分類器に学習させることで, 分類モデルを構築し, テストデータを正しく分類できるかどうかで EC 番号の予測精度が評価される. 精度をより高める予測手法を開発することで, 将来的に未知データに対して正確な EC 番号予測を行うことが目指されている. EC 番号予測手法として, タンパク質配列や物理・化学的特性値, 構造的特徴を用いたものなどがある.

化合物の物理・化学的な特性値を用いた MOLMAP[1]では 55 種類の記述子に対してクラスタリングと Random Forests (RF)[2]を用いて EC 番号の 3 桁目までを予測し, 正解率 0.85 を得ている.

分子の結合関係を 2 値で表すフィンガープリントに着目し, 基質と生成物の部分構造の差異を記述するバイナリベクトルを用いた Differential Reaction Fingerprint (DRFP) [3]では, 分子単体, 分子と 1~2 つ隣に隣接する分子の結合に対する部分構造を基質と生成物ごとに抽出し, 多層パーセプトロンで EC 番号の 3 桁まで予測し, F1-Score として 0.77±0.01 を得ている.

4. 提案手法

本研究では, 有機合成に用いる最適な酵素候補探索に焦点を当て, 210 種類の RDKit 記述子[4]を用いて基質から生成物への変化を表現する. RF と特徴選択, オーバーサンプリングを組み合わせ, EC 番号の 3 桁まで予測するモデルを開発する.

主に構造変化を捉えるフィンガープリントに, 物理・化学的特性値を加えることで, 酵素反応の特徴をより詳細に捉えることが可能になると考えられる. フィンガープリントの次元を考慮し, RDKit の 125 種類の物理・化学特性値とフィンガープリント表現に類似する 85 種類の部分構造のバイナリ値を用いることで, 学習コストを抑えつつ, 酵素反応変化の特徴をより表現できるモデル作成が期待できる.

データ作成では, EC 番号が割り当てられた酵素反応を扱う 4 つのデータベースから得られる, 化学反応を文字列化したデータセット[3]を加工し, 特性値 210 種類の基質から生成物の変化を計算し, 酵素反応を 210 次元の特徴ベクトルで表現する. データ加工を行い, EC 番号予測モデルの構築と評価には 47090 種の特徴ベクトルを使用する. モデル構築では分類精度向上を目的として, 必要な記述子だけを用いるための記述子選択を行う. EC 番号の分類粒度を考慮し, 1 桁目分類, EC X (X=1,2, … , 6)の 2,3 桁目分類に分割し, それぞれで分類精度を高める記述子を選択する.

また, 酵素反応データは不均衡なデータとなっており，EC 番号の多数クラスに比べ少数クラスの正分類が難しい, そのため閾値 T 個以下のクラスを T 個まで，SMOTE[5]を用いてオーバーサンプリングを行う.

初めに，データを 4:1 で学習データとテストデータに分け, 学習データに対して層化 5 分割交差検証を用いる. 各分割で検証用学習データと検証用テストデータに分割し, 検証用テストデータに対して SMOTE を適用する. 閾値は各分類のデータ合計数に応じて決定した. なお，学習時間を考慮し, データ数の多い EC 1 桁目分類には SMOTE は適用しない. 最終的に 7 回分の記述子組合せをマージし, 重複を削除したものを選択された記述子として用いる.

次に，1~3 桁目分類における RF のパラメータ調整を行う. 決定木数と最大深さの組合せに対してグリッドサーチを適用し, 記述子選択と同様に各分割の検証用学習データに SMOTE を適用する. 評価値が高い組合せを最適モデルとして用い, テストデータに対する多クラス分類を行う. 各クラスに対する Precision, Recall, Macro F1-Score，および全体の Accuracy でモデルの分類精度を評価する.

5. 数値実験ならびに考察

本研究では，予備実験と本実験を行う．予備実験では，本実験のための記述子選択を行い．本実験では, グリッドサーチによって最適モデルを作成し，予測モデルの分類精度を評価する.

予備実験の結果では，93 種の記述子が選択された．本実験の結果では，決定木数 300，最大深さ 90 が得られ，Accuracy 0.95，Macro F1-Score の全体平均値が 0.79 となった．表 1 に本実験結果を示す.

表 1. 本実験結果

ECクラス	データ数	Precision	Recall	F1-Score
EC 1.X.X	1601	0.80	0.78	0.78
EC 2.X.X	5789	0.83	0.81	0.81
EC 3.X.X	1345	0.81	0.87	0.83
EC 4.X.X	462	0.86	0.85	0.84
EC 5.X.X	67	0.66	0.71	0.68
EC 6.X.X	154	0.96	0.75	0.81
合計	9418			
Macro Average		0.81	0.80	0.79
Weighted Average		0.96	0.95	0.95
Accuracy		0.95		

DRFP を用いた手法と同程度の評価値が得られたが, 記述子選択とパラメータ調整に多くの時間がかかる点, データ削除によって元のデータセットよりデータ数が少なかったことなどから, 利用データ，学習手法の改善が望まれる．また，各 7 回で選択された記述子の特徴を分析し，類似する記述子をまとめることで次元削減を行う. もしくは，記述子の重要度を評価して，記述子に重みづけを行う方法などが分類精度向上に有効であると考えられる.

6. おわりに

本研究では，物理・化学的特性値とフィンガープリントによる予測手法を組み合わせ, 化学反応に最適な酵素候補を EC 番号として予測する手法を開発した．今後の課題として，分類粒度がより細かい EC 番号の 4 桁目まで予測する手法の開発が挙げられる．また，実際に提案モデルを有機合成の研究者に利用してもらい, モデルに対して現実的な実験条件や予測誤差のフィードバックを行うことが挙げられる.

参考文献

[1] D. A. Latino, J. Aires-de-Sousa, “Assignment of EC numbers to enzymatic reactions with MOLMAP reaction descriptors and random forests”, Journal of chemical information and modeling, Vol. 49, No. 7, pp. 1839-1846, 2009.

[2] L. Breiman, “Random Forests”, Machine Learning, Vol. 45, pp. 5-32, 2001.

[3] D. Probst, “An explainability framework for deep learning on chemical reactions exemplified by enzyme-catalysed reaction classification”, Journal of Cheminformatics, Vol. 15, No. 113, 2023.

[4] “The RDKit Documentation”, https://www.rdkit.org/docs/GettingStartedInPython.html#list-of-available-descriptors, 閲覧日 2024. 1. 31.

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, “SMOTE: synthetic minority over-sampling technique”, Journal of artificial intelligence research, Vol. 16, pp. 321-357, 2002.