ISS-SP-003

グラフ構造に着目した分子特性の認識

石田　聖†　　宮崎　智††　菅谷　至寛††　　大町　真一郎††
† 東北大学工学部電気情報物理工学科　　†† 東北大学工学研究科通信工学専攻

1. はじめに

近年、ニューラルネットワークを用いた分子特性の認識に関する研究が広く行われている。他方、グラフ構造に着目した特徴量は考慮されていない。そこで本研究ではニューラルネットワークとグラフ構造の特徴量を組み合わせることで、高精度に分子特性の認識をすることを目指す。

2. 関連手法

2.1 特徴量設計　特性認識を行う際、まず分子を特徴量に置き換える。この特徴量をもとにコンピュータに学習させ、特性認識を行わせるのだが、この特徴量の設計方法によって認識の精度は大きく変わる。本研究では分子のグラフ構造に着目した以下 2 つの特徴抽出手法を参考にした。両手法とも、化合物中の原子をノード、結合をエッジとすることで、化合物をグラフとして扱う。

2.2 Graph Convolution[1]　化合物をグラフとみなし、ノードごとに特徴量を割り当てる。次に目的ノードの特徴を隣接ノード、結合の情報から更新する。この操作を繰り返すことで広域のノードを考慮した特徴量を得る。最終的に全ノードの特徴量をまとめ、化合物の特徴量を得る。ノード間のベクトルの距離、結合次数によって重みをつけることで特徴量の学習を可能にした手法である。

2.3 ECFP（Extended Connectivity Fingerprints）[2]　Fingerprint とは原子の構造的特徴の有無を 0/1 ベクトルで表現したものである。ECFP では、まず固定長の 0 ベクトルと引数を取り一つの整数を返すハッシュ関数を用意する。次にグラフ化した化合物のノードごとにハッシュ値を割り当てる。それぞれのノードのハッシュ値を隣接ノードとともにハッシュ関数に入力し得られる値を新しいノードのハッシュ値とする。一方で得られたハッシュ値に対応するビットに1を立てる。これを繰り返すことで様々なスケールの構造特徴を取得する手法である。

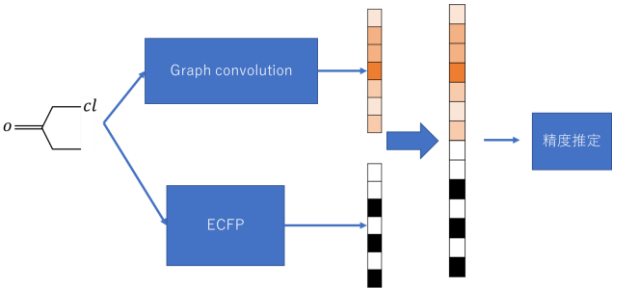

図 1. 提案手法

3. 提案手法

Graph Convolution では原子や隣接原子の情報が重要視され、構造的特徴が欠落していると考えられる。そこで、ECFP の特徴量を付け加えることで、構造的特徴を補い精度の良い特性認識を可能にしようと考えた。そこでこの二つの手法で得られた特徴量をつなぎ合わせるという手法を考えた（図1）。

4. 実験・結果

4.1 データセット　本研究では次のデータセットを用いた。
BACE,SIDER,HIV,BBBP,ClinTox,ToxCast,Tox21。
生体内の化学反応において、人体に影響があるかないかを化合物ごとに判定する。

4.2 評価手法　今回は ROC-AUC score を用いて評価を行った。データを正、負の 2 クラスに分類するタスクにおいて、負のデータを間違って正と判断した確率を偽陽性率、正のデータを正しく正であると判断した確率を真陽性率という。偽陽性率が低いとき、真陽性が高いモデルほど良いモデルといえる。ROC-AUC score ではこのようなモデルほどスコアが高くなうように計算する手法である。

4.3 結果　ECFP, Graph Convolution を提案手法のスコアと比較したところ、その他二つの手法に比べて平均的に提案手法のスコアが良いという結果であった。

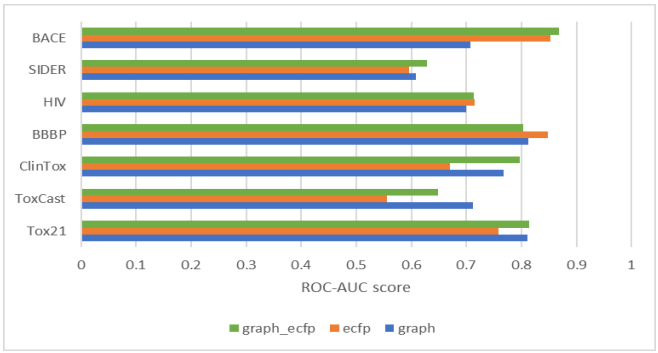

図 2. 認識精度の比較

5. 今後の課題

今後はECFP の特徴量を学習させ、さらなる精度向上を図る予定である。

参考文献

[1]Han Altae-Tran, Bharath Ramsundar, Aneesh S. Pappu, and Vijay Pande, Low Data Drug Discovery with One-shot Learning 2016

[2] Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. J. Chem. Inf. Model. 2010, 50(5): 742-754