分子のグラフ表現と機械学習

瀧川一学(北海道大学・化学反応創成研究拠点)

機械学習分野のスター研究者 Andrew Ng 教授は「“Applied machine learning” is basically feature engineering.」と表現した。現在、身近な ICT 社会基盤への機械学習技術の浸透が急速なペースで進んでいるが、機械学習にどのような特徴量(feature)を入力するのが良いかという根本的な難点は残ったままである。特徴ベクトル表現が決まれば既存の多様な機械学習手法が適用できるが、実際はこの特徴生成・特徴設計のステップが最も困難で、多角的専門知識や手間・時間を要し、かつ、最も予測精度を左右する。画像や音声など定型の生計測が自明に多量に得られる場合には深層学習による特徴表現学習が極めて有望であるものの、一般には特徴表現設計は feature engineering と呼ばれる人手の職人的な試行錯誤に依ってしまっている。これでは結局対象問題の特性を「機械が学習」しているというより「機械学習ユーザが学習」している状態であり、技術的解決が最も期待されている課題の一つと言える。

化学で機械学習を使う場合、分子をどのような特徴量で表現すれば良いかという問題となる。実際には分子の構造は単結合での回転や孤立電子対を持つ原子での立体反転の空間的自由度があり、相互変換可能なこうした多数の立体配座異性体の総体として現れると考えられる。特に生物活性、毒性や医薬品の薬物動態など多分子との相互作用で定まる性質をモデル化する場合、こうした側面をどのように考慮するかは依然として難しい問題である。

分子グラフはこのような場合に長らく使われてきた分子表現の方法である。原子を「点」、原子間の結合を「線」で抽象的にモデル化し、分子の構造は「点(頂点)」を「線(辺)」でつないだネットワーク(グラフ構造)で表現される。化学構造は原子間の結合の組み替えの組合せパターンであり、化合物ライブラリの構造検索やデータ解析にも活用されてきた。機械学習で活用する場合には、頂点や辺にそれぞれ原子や結合に関する多変量の特性値(潜在変数ベクトル)を付与し、構造情報と構成原子や結合の特性量から構成要素の組合せパターンを捉えることでその分子の関心特性値をモデル化する。

本発表ではこのような分子のグラフ表現とそれに基づく機械学習の概要を簡潔に紹介する。近年、こうしたグラフ表現された対象を直接入力にとるニューラルネットワークモデル(Graph Neural Networks; GNNs)が多数発表され、分子活性や分子物性の予測モデリングだけに留まらず、幅広い機械学習タスクで利活用が研究されており極めてホットな研究トピックとなっている。一方で、画像・音声・テキストといった対象で深層学習の高い有用性が確認されてきた事実と比べると、未だに GNN は発展途上の技術であり、様々な技術的問題や限界もよく議論されてきた。GNN は “deep(深層)”にしても有効性が確認できない(over-smoothing、over-squashing 等の問題が起きる)、画像(BiT や SimCLRv2)や言語(BERT や GPT-3)のような大規模データを活かす有効な転移学習の方法や有効性が確立していない、などの技術的関心の他にも、GNN の予測精度の検証の方法自体への注意や、データそのもののバイアスなど様々な未解決な問題も含めて、分子のグラフ表現研究の現在を知って頂く機会としたい。

PROFILE

瀧川一学(北海道大学・化学反応創成研究拠点)

2004 年北海道大学大学院工学研究科博士後期課程終了。博士（工学)。北海道大学大学院情報科学研究科 博士研究員(COE)、京都大学化学研究所バイオインフォマティクスセンター助教（大学院薬学研究科 助教兼務)、北海道大学大学院情報科学研究科 准教授、JST さきがけ研究員などを経て、2019 年より理化学研究所革新知能統合研究センター iPS 細胞連携医学的リスク回避チーム研究員、および、北海道大学化学反応創成研究拠点(WPI-ICReDD)特任准教授(クロスアポイント)。専門は離散構造を伴う機械学習および自然科学のデータ集約型研究。IEEE Senior Member。