6. AIを用いた創薬の実際と今後の展望
──化学構造のための深層学習技術GCNの紹介とその応用

小島 諒介／奥野 恭史 京都大学大学院医学研究科ビッグデータ医科学分野

医療・創薬において深層学習を用いた新たな手法が日々考案されており，画像認識で大きな成功を収めた畳み込みニューラルネットワーク（convolutional neural network：CNN）は医療における画像診断などに応用され，従来法を大きく上回る性能を発揮している例も少なくない。特に，経験的に設計された低次元の特徴量表現に変換することなく，直接，画像データを入力として必要な出力（例えば，ラベル）を得るためのニューラルネットワークを学習させるという方法論はend-to-endアプローチと呼ばれ，深層学習の基本的な考え方の一つになっている。

このend-to-endアプローチは，特徴量に変換する際の情報の欠損を最小限にとどめることができ，システム全体の最適化も容易であることから，医療・創薬における多くのタスクにおいても，性能を出す上で重要な考え方となっている。しかし，医療・創薬では，医用画像のほかにも，臨床データが保存されている電子カルテなど，専門分野の知見やエビデンスを集約した種々のデータベース，さらには化合物やタンパク質の分子構造データに至るまでの多岐にわたるデータを扱う必要がある。しかし，これらのデータは，必ずしもCNNやそれ以前から使われていた多層パーセプトロン（multi-layer perceptron：MLP）などのニューラルネットに情報の欠損なく入力できるようなデータ構造を持っているわけではない。本稿では，これらの従来の深層学習では扱うことの難しい構造のデータに対する新たな技術として，グラフ構造を扱うニューラルネットワークであるグラフ畳み込みネットワーク（graph convolutional network：GCN）を紹介する。また，GCNを医療・創薬に応用していく上で，この技術がどのような位置づけにあり，そのほかの関連技術と合わせてどのように活用していくかについて述べる。本稿では，まず，創薬におけるグラフ構造を扱う問題を例として挙げつつ，GCNに関する技術的な紹介を行う。その後に，医学研究におけるわれわれの取り組みに関連したGCNの活用を紹介し，最後にまとめと今後の展望について述べる。

創薬におけるグラフ構造

創薬では，医薬品の種となる候補化合物の数が膨大であることから，すでにある実験結果から計算機を用いて新たな化合物の活性や特性を予測するということが重要な課題である。具体的には，化合物やターゲットとなるタンパク質の性質を既存の実験データから予測する問題や，タンパク質と化合物との間の反応・活性を予測するといった問題が代表的である。

こういった問題を扱うためには，入力を化合物やタンパク質とし，出力を活性の有無といった機械学習のモデルを構築する識別問題に落とし込むことが一般的である。化合物やタンパク質の情報を機械学習モデルに入力するために，従来，化合物やタンパク質の記述子を設計し，それらを特徴量表現として用いることが一般的であった。例えば，化合物の場合にはextended connectivity fingerprint（以下，ECFP）などの記述子，タンパク質の場合にはPROFEATといった記述子が知られている。しかし，これらの記述子の多くは非可逆変換であり，元のデータの情報を完全に保存することが難しく，そのため，タスクに応じて適切な記述子を選択する必要がある。例えば，従来の記述子の中でよく用いられるECFPは，化合物構造式の部分構造の有無を数値化したベクトルで表現したものである。ECFPは，部分構造の大きさやベクトルの大きさについてはあらかじめ決めておく必要があり，部分構造の有無のみを表現しているために，「その部分構造が全体のどこにあるか」といった化合物の全体の構造に関する情報は失われてしまう非可逆の変換である。化合物構造式に対してend-to-endアプローチをとるためには，化合物構造式を直接入力できるようにする必要があり，そのための表現方法の一つが化学構造式を「グラフ」とみなす方法である。

ここでいう「グラフ」とは，ノード（点）とエッジ（辺）から構成されるデータ構造であり，実用上ではこれらのノードやエッジには「属性」が付与されていると考える。このようなデータ構造を考えると，例えば，化合物の構造の場合には原子はノードとみなし，結合はエッジとみなすことができる。この時，原子タイプなどの情報は属性としてノードに付与し，結合の種類はエッジの属性に付与することができる。グラフのエッジはノード間のつながりのみを表現してお

〈0913-8919/18/￥300/ 論文 /JCOPY〉