深層学習を用いたタンパク質会合系小角散乱データの成分解析

藤澤　哲郎
岐阜大学・工学部

人工知能技術は新しい技術であり非常に高いポテンシャルを有している。しかし、その応用はほとんどが画像や音声データ解析に集中している。人工知能のアルゴリズムはたくさんあるが、そのアルゴリズムからわかることわからないこと、機械学習に必要なデータの要件、分析精度などは当該分野の研究者が経験を蓄積していかなくてはならない。2018年には溶液散乱の分野でも機械学習により、散乱から得られるパラメータを基に散乱体の形を導出する例が報告されている[1]。しかし、この研究では、スペクトルデータそのものではなく、それから得られたパラメータに対して機械学習を適用している。一般に、小角散乱も含めて一次元のスペクトルデータに対して機械学習の適用例がほとんどない。

本研究では、「複数の成分を含む混成系の散乱曲線から構成成分の組成分率や会合数を求める」という問題に人工知能技術を適用し検証してみた。従来、このような波形分離には多変量解析を応用した最小自乗法（MCR 法）が用いられてきたがあいまいさが常につきまとってきた[2]。2017年には、MCR 法を溶液散乱スペクトルに適応させた COSMICS 法が提唱されたが依然あいまいさは存在し、しかも非公開である[3]。全く新しい原理による波形分離法が確立されれば従来の方法とあいまって、あいまいさは大幅に改善されることが期待される。本研究においては人工知能技術の中でも、特に深層学習に注目した。深層学習は人間の神経回路網を模倣した機械学習手法の一手法で、与えられた教師付きデータ（訓練データ）から特徴を自動的に検出し学習を行っていく。深層学習は画像、音声などの認識や識別問題に対し高い正解率を誇っているため、この識別能力を小角散乱データ解析に応用できることをシミュレーションデータにより検討した。

本研究で最も重要な点は、いかにコンピュータが得意なデータ構造に散乱データを適合させるかである。散乱曲線は散乱ベクトル q と散乱強度 $I(q)$ で表現されており、隣接する各 $I(q)$ 値は非常に高い相関を有する。この点は、音声や文章などの時系列データと共通とみなし、時系列で変化するデータに特化した再帰型ニューラルネットワークを実装した。

学習するデータは組成分率が異なる混成系の理論散乱曲線 q-$I(q)$であり、それぞれ会合数、組成分率などの教師データをセットで持っている。学習後、組成等を予測する場合は散乱曲線 q-$I(q)$のみを入力とし、予測の結果が会合数や組成分率として出力される。
本研究ではこの教師データを体積分率、z 分率、会合体が存在するかの有無の 3 種類用意し、学習、検証を行った。正解率はそれぞれ 87%、87% 、95%に及んだ[4]。

[1] Franke D. (2018) Biophys. J., 114, 2485.
[2] Tauler, R. (1995) Chemom. Intell.Lab.Syst.,30,133.
[3] Herranz-Trillo, F. et al. (2018) Structure, 8, 7204.
[4] 鈴木皇陽、平成 29 年度岐阜大学大学院工学研究科修士論文