ISS-A-053

化合物データ増強へのGAMOの応用に関する考察

鐘川　凌　　福田　龍樹
北九州工業高等専門学校　専攻科

1. はじめに

新材料開発は今日まで活発に行われてきた研究の一つである. しかし、従来までの新材料開発手法は研究者個人の経験や勘などを頼りに材料の合成実験とその合成物質の特性の解析を試行錯誤的に繰り返すという効率が悪いものである. これは、組成構造が複雑であり、目的とする構造の合成方法の検討をつけることが困難であることが原因であり, この問題の解決には目的とする構造や特性などのパラメータを逆算的に特定する手法の開発が必要不可欠である. その方法として近年マテリアルズインフォマティクス(以下 MI)と呼ばれる材料科学とデータ化学とを組み合わせて新材料開発を行う手法が注目されている.

しかし、MI にも課題点があり、主な問題点としては材料データの不足が挙げられる. 材料データは企業や研究室によって秘匿されているものがほとんどであるため, 材料データが大抵の場合不均衡になってしまい, 機械学習などの適用を困難とさせている [1].

そこで, 本研究では少ないデータや不均衡データからの効率的な解析手法の開発を目的とし, その方法としてGAMOと呼ばれるデータ増強モデルを応用する手法を提案する.

2. GAN

機械学習の基本となる概念であるニューラルネットワークを応用したモデルに GAN(Generative Adversarial Networks)がある. このモデルはニューラルネットワークによって構成される生成器と識別器の二つのモデルを用いて学習を行う. 生成器はデータを生成し, 識別器は入力されたデータがデータセット内に存在するものか, 生成器が生成したものかどうかを識別する. この流れを二つのモデル間で繰り返し行うことで, 生成器が最終的によりデータセット内のデータに近い特徴量をもつデータの生成するようになる. しかし, GAN には学習が不安定になりやすいことや同じようなデータばかりを繰り返し生成する現象が起こることがあるため, データ増強手段には適していない. そこで, より効率的なデータ増強手法として提案されたのが GAMO である.

3. GAMO

不均衡データへの対処方法としてデータ増強という手法がある. この手法は元のデータにノイズなどを加えることで一つのデータを複数のデータとみなすものであるが, GAMO は GAN のアイデアを応用してデータ増強を行う. GAMO では GAN と同様にデータを生成する生成器とデータを識別する識別器を使用するが, さらにデータの特徴量を分類する分類器を加えた合計三つのモデルを互いに学習させ合い, データの増強を行う.

また, GAMO には凸結合という計算方法を用いて特徴量を生成するという特徴がある. 特徴ベクトルに重みをかけて足し合わせる計算であり, 重みのみを足し合わせたときの合計が1である制約を持つ. この計算により元のデータから大きく離れすぎない新しい特徴量の生成を可能とする.

$$\alpha_1 x, + \alpha_2 x_2 + \cdots + \alpha_n x_n \quad (1)$$

4. 現在の進捗

本研究ではデータの前処理をした後に GAMO を利用してデータの増強を行うが, 現段階ではデータの下処理までが完了している. 化合物データには Python ライブラリである DeepChem で公開されている delaney-processed データセットを使用した. このデータセットは1128種類の化合物についての特性についてまとめられたデータセットである. 本研究では, まず化合物データを機械学習などが行える状態に処理するためにベクトル化を行う. ベクトル化には Morgan フィンガープリントと呼ばれるものを使用した.

また, GAMO は複数のクラスが不均衡であることを想定したものであるため, 化合物データを複数のクラスに分類する必要がある. そこで, 本研究ではクラスタリングという手法を用いて化合物データを五つのクラスに分類した.

5. 今後の展望

今後は前処理したベクトルを GAMO へ入力し, データ増強をすることを目指す. また, GAMO は画像データ増強のためのモデルであるため, そのモデル構造を化合物データ用に調整することも今後の課題である.

参考文献

[1] 伊藤 聡, “日本のマテリアルズインフォマティクス研究, ” 人工知能学会誌,vol.34 no.3,pp.325-329, May 2019.

[2] Sankha Subhra Mullick, Shounak Datta, Swagatam Das, “Generative Adversarial Minority Oversampling” https://arxiv.org/abs/1903.09730, Submitted on 22 Mar 2019.