ベイズ推定

統計や機械学習における確率を用いた計算方法の一つに「ベイズ推定」がある。ベイズ推定とは、観測された事象から原因である事象の確率を推定する手法であり、「ベイズの定理」を利用するものである。具体例としては、ベイズ推定を利用した「ベイジアンフィルター」[1]が挙げられる。ベイジアンフィルターは、スパムメールに多く含まれることが分かっている単語X（これはスパムメールと非スパムメールにおける単語Xの出現頻度を比較することで求めることができる）に着目してフィルター処理を行う。ベイズの定理を使うと、単語Xを含むメールがスパムである確率（単語Xという事象が起こったあと、メールがスパムである確率）を、スパムメールの割合（受信したメールのうちスパムメールだったものの割合）、非スパムメールの割合、スパムメールに単語Xが含まれる確率、非スパムメールに単語Xが含まれる確率から計算することができる。

アンサンブル学習による高精度の予測

精度の高いモデルを手軽に構築できる手法として注目されているものの一つに「アンサンブル学習」がある。アンサンブル学習とは、複数のモデル（学習器）を組み合わせて一つの学習モデルを生成する手法である。アンサンブルを構成する個々のモデルは必ずしも高い精度を持たない「弱学習器」でも良いため、開発が容易になる※31。Kaggle※32のコンペなどでも上位ランクを占めるものの多くがアンサンブル学習を利用したものである※33。アンサンブル学習の手法の一つにランダムフォレストがある。これは、弱学習器として決定木を使うもので、複数の決定木の結果から、分類の場合は、各々の決定木の分類結果の多数決をとり、回帰の場合は、各々の決定木の予測結果の平均を計算するなどして、結果を得る。

グラフ構造データによる予測（GNN；Graph Neural Network）

ディープラーニングでは、画像や映像など独立したデータの集合を学習データとしているが、最近では、グラフ構造（データ間のつながりを持った構造）のデータを入力とするニューラルネットワーク、GNN（Graph Neural Network [2]）が登場している。GNNでは、グラフ構造を行列表現し、これを入力としたニューラルネットワークを構成する。畳込みニューラルネットワーク（CNN）やオートエンコーダー、再帰型ニューラルネットワーク（RNN）を構築することもできるため、広範囲な応用が可能だ。分子構造の予測や自然言語処理などへの応用がある。例えば、生物が体の中で合成する天然合成物質（A）の合成の起点となるスタート物質（B）（2万件中千程度しか既知でない）を、Aの分子構造をグラフとしたGNNを利用して、予測する※34などがある。

※31 ただし、全体としての性能を上げるためには弱学習器の予測性能に多様性（それぞれが強み弱みを持つ複数の弱学習器を持つこと）を確保することが必要である。
※32 Kaggle: Your Home for Data Science <https://www.kaggle.com/>
※33 Kaggle Ensembling Guide <https://mlwave.com/kaggle-ensembling-guide/>
※34 “化学構造を手掛かりにしたデータサイエンスの手法で、天然物化合物の生合成経路の予測に成功！”<http://www.naist.jp/pressrelease/2019/07/006019.html>

ディープラーニングとシミュレーションを組み合わせた予測

ディープラーニングの応用分野として、シミュレーションやデータマイニングなどのIT技術を使って材料開発、探索を行うマテリアルズ・インフォマティックス（Materials Informatics）や医療分野での薬剤開発、遺伝子解析などのバイオ・インフォマティックス（Bio Informatics）がある。国立研究開発法人日本医療研究開発機構は、薬剤とタンパク質の相互作用を予測する、新たなディープラーニング手法を開発した[3]。これにより薬剤とタンパク質の相互作用部位を特定・可視化、予測結果の妥当性の確認が可能となり、新薬開発が加速するとした。そのほか、地学、地震学、天文学、気象など、これまでいわゆる科学技術計算としてのみコンピューターを使ってきた分野でも、ディープラーニングを使うことで、高速な探索、解析、予測などが可能になってきた例がある。

また、ディープラーニングとシミュレーション技術を組み合わせたものとして「データ同化」がある。シミュレーションにはモデル化に際して現実の現象との乖離があり、なんらかの誤差を潜在的に持つ。データ同化とは、シミュレーション結果と現実の観測値を比較し、観測値に適合するようにシミュレーション結果を補正するための技術である。従来は、統計的な手法が利用されていたが、誤差関数が持つ高次元性や非線形性から所望の性能が得られなかった。これも、深層ニューラルネットで非線形関数を近似することで、シミュレーションと現実の間にある非線形関係を推定できるようになった。これは、演算性能の向上やディープラーニング技術の進歩により、実用段階に入ったといえる。

(3) 最新技術動向

タンパク質構造予測（AlphaFold）

DeepMindの研究チームは、アミノ酸の配列からタンパク質の立体的な形状を予測する「タンパク質構造予測」に対して、従来にないアプローチを採用したAIシステム「AlphaFold」を導入した。同システムの最大の特徴は、言わば「2段構えのAI学習」を実現したことにある。1段目の学習においては、既知の遺伝子配列とタンパク質のデータを学習データとして、アミノ酸が互いに結合するときの結合の長さ（Distance）と角度（Angle）に関する予測モデル（Distance PredictionとAngle Prediction）を構築する。続く2段目の学習において、AlphaFoldの革新性である「予測モデル精度の改善」が実行される。この学習では、まず予測モデルの予測精度をスコア化する関数（Score）を作成し、この関数に対して、1段目の学習で活用した学習データとは別のタンパク質構造に関する情報を与えることで、予測精度を表すスコアを最適化するように学習を実行する。この学習においては「勾配降下法（Gradient Descent）」が適用される。つまり、アミノ酸配列の情報（Protein Sequence）から、アミノ基の間の長さ（Distance）と角度（Angle）を予測（3次元構造を予測することに相当）し、さらに全体の3D構造が妥当かどうかを判断するスコアリング（Score）を組み合わせることで、3次元構造（Structure）を予測している。距離・角度の予測、スコアリングにそれぞれ個別のニューラルネットを利用している（**図2-2-16**）。