医療、自動運転

機械学習モデルによる予測の「自信」を較正する

D3センター
准教授 **早志 英朗**
Researchmap https://researchmap.jp/Hayashi_Hideaki/

研究の概要

医療や自動運転といった信頼性が必要とされる分野で機械学習モデルを応用する際に、モデルがどの程度「自信」を持って予測しているかを参照できれば有用である。クラス識別では、モデルの予測に対する自信の程度を示す指標として、信頼度と呼ばれる予測クラスに対する事後確率がしばしば用いられる。しかし、特に近年の深層学習モデルにおいて、この信頼度と実際の予測正解率が一致しないことが問題であった。本研究では、機械学習モデルが出力する信頼度を較正する手法を提案する。提案法では、深層ニューラルネットワークによる識別モデルの最終層において、生成モデルを同時に学習する。これにより、モデルが入力データの生起確率を考慮してクラス事後確率を算出できるため、より正確な信頼度を獲得できる。

研究の背景と結果

大規模言語モデルや生成 AI の驚異的な発展に伴い、機械学習モデルの信頼性が重要視されるようになった。モデルが高精度でもっともらしい出力をするようになったとしても、人間による最終的な意思決定が必要な場合は多い。その場合、モデルがどの程度自信をもってその結果を出力しているかを人間が参照できると有用である。

機械学習におけるシンプルかつ重要なタスクであるクラス識別では、識別器が予測結果とともに出力する確率が意思決定の参考となる。クラス識別は入力がどのクラスに属するかを予測するタスクであり、例えば画像にどの物体が写っているかを推定する画像認識が該当する。モデルは画像を入力として取り、どの物体クラスが写っているかの確率を出力する。ここで、クラス確率の最大値は信頼度と呼ばれ、予測の信頼性の重要な指標となる。

しかし、機械学習モデルの出力する信頼度はあてにならないケースがしばしばある。「あてにならない」とは、信頼度が実際の正解率と一致しないことを意味する。例えば、学習済みの識別器に多数のデータを入力し、それぞれに対する予測ラベルと信頼度を得たとする。識別器が「信頼度70%」と出力したデータを集めると、正解率は70% になることが期待される。ところが、特に近年の深層学習モデルにおいて信頼度と正解率は一致しないことが多い。

本研究では深層ニューラルネットワークによる識別モデルの最終層において、生成モデルを同時に学習するアプローチにより、学習中に信頼度を較正する手法を提案した。実験では、一般物体画像や医用画像の識別タスクにおいて信頼度較正能力を示した。特に、ラベル付き学習データが限られる半教師あり識別タスクにおいて、expected calibration error と呼ばれる信頼度較正を測る指標を従来法に比べて大幅に改善することを明らかにした。

研究の意義と将来展望

信頼度が機械学習モデルの予測の不確かさを正しく反映していれば、実世界の意思決定において有用である。例えば、医用データの識別を病気の診断に応用するシナリオにおいて、モデルは正常・異常等のクラス予測結果とそれに伴う信頼度を出力する。診断という最終的な意思決定は医師によってなされるため、医師はモデルが出力する信頼度を参考に、予測結果を採用するか否かを決定する。信頼度が一定値以下であれば予測結果を棄却し、改めて目視による精密な診断を行うことができる。

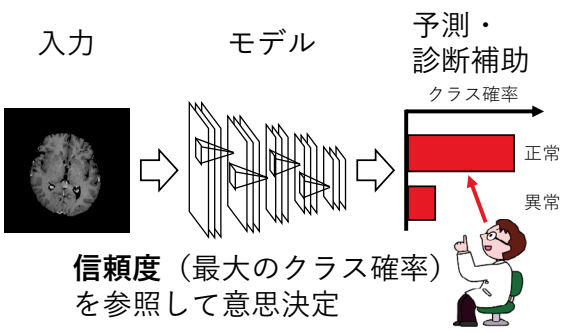

図1

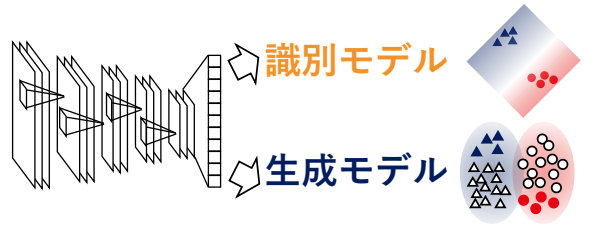

図2

特　　許	
論　　文	Hayashi, Hideaki. A hybrid of generative and discriminative models based on the Gaussian-coupled softmax layer. IEEE Transactions on Neural Networks and Learning Systems. 2024 (early access). doi: 10.1109/TNNLS.2024.3358113
参考URL	https://sites.google.com/view/hideakihayashi/home
キーワード	機械学習、パターン認識、人工知能