外部文書検索による大規模言語モデルの データ予測性能の改善

西川　幸志

近年，データ分析は，膨大な情報から有用な知見を抽出する技術として，多様な分野で意思決定の基盤として活用されている．しかし，データの規模や複雑さが増大するにつれて，専門家による手動の分析では効率が低下し，またデータ分析を担う人材の不足も深刻化している．こうした背景から，データ分析の自動化技術への期待が高まり，とりわけ大規模言語モデル（LLM）を用いた自動化への注目が集まっている．実際，データの前処理や特徴量エンジニアリングなど，近年の LLM 活用によって段階的に自動化されつつある．一方で，LLM が内部に保持している知識だけでは限界があり，特定の領域で高度な専門性が必要とされるタスクを十分にこなすことは難しい．医療分野を例にとれば，心疾患の予測モデル構築においては，「年齢」「アルブミン値」「BNP 値」など，臨床試験を通じてリスク因子としての重要性が確立された指標を重視することで高い予測精度を実現しているケースが報告されている．しかしながら，一般的な LLM は「与えられたデータセットのみ」をもとに学習するため，こうした外部の専門知識をモデルに取り込むことが難しく，結果として重要な特徴量が十分に評価されない可能性がある．特に医療や金融のようにドメイン知識が欠かせない場面では，外部から知識を取り入れる仕組みがないとモデル性能には限界が生じる．そこで本研究では，LLM に外部からのドメイン知識を取り込む手法を提案する．具体的には，まず表形式データの特徴名やタイトルなどをもとに関連文献を検索し，得られたテキスト情報を LLM に与えることで，ドメイン固有の専門知識を事前に付与する．次に，LLM はこれらの知識を踏まえ，各特徴量がターゲット変数とどの程度関連するかを表す「事前重み」を推定する．従来，こうした重みは学習過程で自動的に獲得されるが，本手法では外部知識を用いて訓練前に導出し，これを損失関数の正則化項に組み込むことで，実際の学習時に「外部知識に基づく重要度」と「データによって示唆される重要度」を統合することを狙いとする．これは専門家が特徴量の重要度を見立てる作業に類似しており，モデル構築の初期段階からドメイン知識を組み込むうえで有効なアプローチと期待される．本研究では，医療分野の 103 件の表形式データセットと，約 3,700 万件の生物医学文献から得られる情報を外部知識として用いて，上述の手法が予測精度の向上に寄与するかを検証した．その結果，少量データの環境で外部知識を事前重みに反映させた場合，最終的な学習後の重みが事前重みに近づく傾向が見られ，特徴量の重要度をある程度正しく把握できる可能性が示唆された．しかし，予測精度の向上は常に明確には確認されず，線形モデルの使用や文献検索精度の問題など，知識活用の効果を十分に引き出すにはさらなる工夫が必要であることも分かった．

（指導教員　　加藤　誠）