決定木モデルに有効なデータセットの特性を表す指標の探索

毛戸宇仁

機械学習は，近年の関連技術の発展を受けて大きな注目を集め，また実社会での活用も広まっている．その中でも深層学習などニューラルネットワークを基盤とした手法は多くの分野で活用され，その範囲も今後広がることが期待されている．しかしながら機械学習は結果のみが得られてその理由が分からないため，学習の適切さや結果の妥当さを検討することが難しい．その傾向はニューラルネットワークでは特に顕著であり，現状では意思決定への活用など責任が求められる場面では使いづらいとされる．

この課題に対して，機械学習の結果の理由を説明する追加の情報を学習したモデルから抽出することで，理由についての人間の理解を助ける試みが説明可能 AI と呼ばれる技術や分野である．その中には，複雑なモデルの可読化を目的とした大域的な説明と呼ばれるアプローチがあり，ここでは解釈性の高い機械学習による近似で実現する手法に注目した．本研究では解釈性の高い機械学習の代表的な存在である決定木モデルをとりあげ，その有効範囲についての検討を行う．

決定木モデルの性能を評価した先行研究によると，決定木モデルは高い分類精度を示す場合が存在する一方で，そうなる場合は学習対象のデータセットによって限定されると言われている．このことから決定木モデルが高い分類精度を示す場合と，それ以外の場合との間に存在する差異はどの様なものなのかという問題意識を得た．そのような差異は学習対象のデータセット何らかの特性の違いによると考え，決定木モデルの分類精度の高低に影響を与えるデータセットの特性の探索を目的に実験を行った．

実験では iris データセットを対象に，その状態を様々に変化させることで決定木モデルによる分類精度への影響を調べた．まずデータセットのサイズを特性と考えその縮小を検討したが，分類精度に影響はなかった．次にデータセットの平均を特性と考え，平均が出来るだけ大きく変化する縮小データセットを多数作成した．実験の結果，分類精度が若干落ちるいくつかの縮小データセットが見つかり，それらの標準偏差が iris データセット全体と異なる傾向がみられた．そこで，標準偏差を特性と考え，標準偏差が大きく変化する縮小データセットを複数作成した．実験の結果，標準偏差の変化を大きくした縮小データセットで分類精度の低下は認められなかった．

今回は，データセットのサイズ，平均，標準偏差を特性として実験したが，それらが分類精度に影響を与えないことが分かった．今後は，それ以外の特性について検討していくとともに，より理論的な考究が必要である．

（指導教員　中山伸一）