2C-06

回帰・分類問題における能動学習の研究動向と課題に関する一考察

阪井　優太†　　小林　学†　　後藤　正幸†

早稲田大学†

1. 概要

一般に教師あり学習モデルの学習で用いられる教師ラベルの付与(アノテーション)には多大なコストが発生する．そのような状況において，逐次的に教師ラベルを付与するデータをサンプリングすることでラベルの付与コストの合計を抑えつつ精度の高いモデルを構築する方法として能動学習がある．能動学習は従来分類問題における議論が非常に多かったが，近年では回帰問題での研究事例も増加傾向にある．そこで本稿では，能動学習における回帰・分類問題における問題設定と最近の研究動向をまとめ紹介し，その課題と今後の展望について考察を行う．

2. 能動学習の問題設定

以降では，説明変数を$\boldsymbol{x} \in \mathcal{X}$，教師ラベルを$y \in \mathcal{Y}$とする．また，$\mathcal{Y}$は二値分類問題においては$\{-1,+1\}$，回帰問題においては$\mathbb{R}$となる．説明変数から教師ラベルを予測する関数を$f$とする．また，予測結果の評価を行うための損失関数を$L(\boldsymbol{x},y,f)$とする．分類問題では，0-1 損失$L(\boldsymbol{x},y,f) = 1(y \neq f(\boldsymbol{x}))$を用いることが多い．回帰問題では，二乗誤差損失$L(\boldsymbol{x},y,f) = (y - f(\boldsymbol{x}))^2$を用いることが多い．真の関数と予測する関数$f$の誤差を最小化することが目的である．

3. 分類問題における問題設定

分類問題における能動学習は従来から多くの問題設定において活用されており，大まかに3つのアプローチに分類される．まず，この分類問題において最も多い能動学習の種類として Pool-based Sampling がある[1]．この問題設定ではあらかじめ大量に目的変数を持たない説明変数のみのデータが蓄積されており，その中から教師ラベルである目的変数の値を付与することで分類精度を向上させることのできるようなデータを選択し学習データに追加する．これを繰り返すことで逐次的に教師データを増やし，分類器の精度を向上させる．過去に蓄積されたデータが大量に存在している状況を想定しており，構文解析や医療画像分類などのラベル付与のコストが非常に高いタスクにおいて活用されている[2]．また，一般的な画像分類データセットを活用した応用研究も多い．

2つ目は，Stream-Based Selective Sampling というアプローチである[3]．この問題設定では，データが逐次的に流入することを想定しており，流れてきたデータに教師ラベルを付与するか否かを決定し，効率的にデータを選択して分類精度の高い分類器を作成する手法である．さらに，この設定を拡張し，次にデータを取得したい特徴量を決定し探索することを可能とする手法が提案されている[4]．

最後は，データ自体を生成しつつラベルを付与する Membership Query Synthes というアプローチである[5]．この問題設定では，次に教師データを得たい特徴量のデータを自身のアルゴリズム内で生成し学習する手法である．教師データを得たい特徴量を適切に生成することができればかなり有用であるが，生成されるデータが人間にとって意味のあるデータとならないことがあるといった欠点が存在する．そのため，分類対象とするドメインのデータの特徴量空間におけるデータの分布も考慮した方法である必要がある．

また，逐次的に分類器の精度を向上させるために選択するサンプルデータの選択方法についても盛んに研究が行われてきた．エントロピーを基準とした手法[6]，識別境界付近のデータを選択する手法[7]などが提案されている．

4. 回帰問題における問題設定

能動学習における回帰問題は近年特に注目されている．逐次的に実験を行うことでデータを効率的に収集し，入力とする特徴量と出力の関係性を正しく推定したいという問題設定が存在する．実応用では，マテリアルインフォマティクスやクリギングなどにおいてこれらの問題設定が活用されている．また，過去に大量に蓄積されたデータを活用した Pool-based Sampling の設定では，画像データを用いた設定が多く，例として動物や人間の顔画像からの年齢推定タスクなどが存在する．

5. 近年扱われる研究課題

ここでは能動学習の課題について述べる．初期に収集したデータがバイアスを含んでいる場

A Study on Active Learning Research Trends and Issues in Regression / Classification Problems
† Waseda University

合や，アノテーションデータが正確でない場合などに良い性能を発揮しないという課題がある．そのため，能動学習では初期データにバイアスがないある程度の数量かつ精度の高い教師ラベルを持つデータを必要としていた．この課題に対して，初期に教師ラベルを持つデータが存在しない場合の学習方法が提案されている[8]．しかしながらこの方法はPool-based Samplingの設定を仮定したものであるため，初期に特徴量のみ存在するデータがない状況からも逐次的に教師ラベルを付与するデータを収集しつつ関数の精度を高める手法の開発が今後望まれる．この分野の研究に受動学習という特徴量空間のみからデータの不確実性を評価する手法が有効と考えられる．

また，能動学習の目的は教師ラベルを付与するコストの削減である．しかしながら近年の問題設定では過去に存在するデータにラベルを付与するコストだけでなく，実験によりデータを収集する際にかかるコストについても考慮した問題設定が研究されている[9]．また，教師データの追加をどこまで行うかという最適な学習停止タイミングを決定する問題も大きな課題として挙げられる．ここで能動学習ではラベルを付与するデータ数を少なくしたいのでクロスバリデーションなどの方法は利用できない．そのためEarly Stoppingが用いられることが多い．ただし，理論的な背景として決定的な方法論が確立されていない．さらに近年盛んに行われている議論として，タスクに依存しない目的関数の設計を強化学習により達成することが挙げられる[10]．従来から課題となっていたデータ分布に適合したデータ選択のための獲得関数を学習対象とするアプローチであり，今後も研究の発展が期待される分野である．

近年の能動学習の研究が画像データに重点を置かれている．しかしながら，画像データに対する適用が問題設定として適切なのかはよく検討するべきである．特にサンプルサイズは大きいが，1つのアノテーションコストが小さい画像データが蓄積されている場合は表現学習や事前学習などほかのアプローチでも高い精度の予測器を作成できる．そのため，能動学習を用いるべき設定であるかはよく確認するべきである．

6. 能動学習との関連分野

能動学習と類似した問題設定として半教師あり学習がよく挙げられる．この半教師あり学習を能動学習と組み合わせることでラベルのないデータの情報も考慮した教師ラベル付与データの選択が可能となる．また，ベイズ最適化とも非常に関係性が深い．ベイズ最適化は何かしらの評価関数を最大化することを目標とするのに対して，能動学習では，予測器の推定精度を高めることに重点が置かれている．そのため，適用する問題設定が異なり，能動学習の方が関数全体の予測結果に対する興味が強いと言える．

7. まとめ

本稿において能動学習の一般的な問題設定をまとめ，今後の研究に向けた課題や周辺分野との関連性について考察した．今後，能動学習を逐次実験計画に組み込んだ研究が必要になると考えられる．その際に発生する初期サンプルの決定方法やデータのバイアスを考慮した手法の開発が期待される．

謝辞

本研究では，日本学術振興会科学研究費基盤No/21H04600(No.19K04914)，JST次世代研究者挑戦的研究プログラムJPMJSP2128の助成を受けたものです．

参考文献

[1] Lewis, D.D., and Gale, W.A. "A sequential algorithm for training text classifiers." *SIGIR'94*. Springer, London, (1994).

[2] Settles, B. "Active learning literature survey." (2009).

[3] Atlas, L.E., Cohn, D.A. and Ladner, R.E. "Training connectionist networks with queries and selective sampling." *Advances in neural information processing systems*. (1990).

[4] Loy, C. C., Hospedales, T. M., Xiang, T., and Gong, S. "Stream-based joint exploration-exploitation active learning." *2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE*, (2012): pp.1560-1567.

[5] Angluin, D. "Queries and concept learning." *Machine learning 2.4* (1988): pp.319-342.

[6] Shannon, Claude Elwood. "A mathematical theory of communication." *The Bell system technical journal 27.3* (1948): pp.379-423.

[7] Scheffer, T., Decomain, C. and Wrobel, S. "Active hidden markov models for information extraction." *International Symposium on Intelligent Data Analysis*. Springer, Berlin, Heidelberg, (2001).

[8] Wu, D., Lawhern, V. J., Gordon, S., Lance, B. J. and Lin, C. T. "Offline EEG-based driver drowsiness estimation using enhanced batch-mode active learning (EBMAL) for regression." *2016 IEEE International Conference on Systems*, Man, and Cybernetics (SMC). IEEE, (2016).

[9] Ueno, T., Ishibashi, H., Hino, H. and Ono, K. "Automated stopping criterion for spectral measurements with active learning. " *NPJ Computational Materials*, (2021).

[10] Woodward, Mark, and Chelsea Finn. "Active one-shot learning." arXiv preprint arXiv:1702.06559 (2017).