能動学習を用いたストリームデータの回帰分類

山本　倫也†　　三浦　孝夫†
†法政大学理工学部創生科学科

1. 回帰分類と能動学習

分類器とはデータをあらかじめ定められた基準(クラス)のいずれかに仕分けるアルゴリズムである．分類器の構築は過去のデータ(学習データ)から特徴を抽出して判定に用いる．しかし次第に当てはまり状況が悪化する．高精度で分類し続けるには，能動学習により改善に有用なデータから自律的に選び、自己改善が必要となる。

本研究では，回帰分類を用いて能動学習によりストリームデータの高精度分類手法を提案する[1]。回帰分類は回帰分析を用いて分類する手法である．回帰分類は，テストデータがどの回帰によりうまく記述されるかによって分類する。テストデータとのユークリッド距離を用いる．テストデータから回帰直線との距離を求めれば、これが当てはまりを示さず，回帰の良さは距離では判断できないため決定係数で重みつける．

回帰分類の問題点は決定係数に大きく依存することである．共に小さな決定係数のとき，当てはまりを定量的に表すとは言えない。更に，ストリームデータではギャップが広がるため、精度が更に悪化する．

能動学習は、分類精度を向上させると判断されたとき学習データに加える機構であり，この選択基準が問題である。回帰分類では当初の学習データを使用し続けるが，提案手法では決定係数を向上させるときに再計算を行う．

2 回帰分類実験

能動学習を用いた回帰分類の有効性を示すため、２地点をクラスとみなし、気象データを回帰分析により分類して測定地点を推定する。実験には気象庁の 1999 年の気象データを用いる．実験対象は 2 地点(クラス)の 1 月の気象データ 31 件とする．目的変数は平均気温，説明変数は平均湿度と日照時間とする．テストデータには実験対象の 2 地点の 2 月気象データを用いる．

第 1 実験は，1999 年１月の東京と京都の気象データをそれぞれ学習データで同年の２月の東京の気象データをテストデータとして用いる．能動学習を用いた回帰分類の正解率は全 28 件中正解数 21 件で正解率は 75%となる．テストデータである 2 月の京都の気象データを能動学習すると，東京 1 月の回帰式の決定係数は 0.2131 から 0.3360 まで増加する．京都の 1 月の決定係数は 0.3641 から 0.3757 まで増加する．**またテストデータである 2 月の京都の気象データを学習したとき東京 1 月の回帰式の決定係数は 0.2131 から 0.3360 まで増加する．京都の 1 月の決定係数は 0.3641 から 0.3757 まで増加する．**

第 2 の実験は，1999 年１月の京都と横浜の気象データをそれぞれ学習データで同年の２月の横浜の気象データをテストデータとして用いる．能動学習を用いた回帰分類の正解率は全 28 件中正解数 12 件で正解率は 42.9%である． 2 月の横浜の気象データを学習したとき横浜 1 月の回帰式の決定係数は 0.1975 から 0.2724 まで増加し,京都の 1 月の決定係数は 0.3641 から 0.4201 まで増加する．

最後の実験では，1999 年１月の東京と京都の気象データをそれぞれ学習データで同年の２月の京都の気象データをテストデータとして用いる．能動学習を用いた回帰分類の正解率は全 28 件中正解数 23 件で正解率は 82.1%となる．2 月の京都の気象データを学習したとき東京 1 月の回帰式の決定係数は 0.2131 から 0.2799 まで増加し、京都の 1 月の決定係数は 0.3641 から 0.4014 まで増加する．

5 考察と結論

決定係数は分類の性能に大きく関わり，初期決定係数が大きいときは回帰分類は十分な精度を保てた．初期決定係数が低い場合は精度が低下するが，能動学習を用いた回帰分類では決定係数が高い水準で上昇し，高い精度で分類可能であることを示した．ただ学習効果が低いときは改善結果が見られず，選定条件の見直しが必要である．

参考文献

[1] Dasgupta, S.: Two Faces of Active Learning, TCS412-19, 2011