アンサンブル学習を用いた文書分類

54234　宮下 塁
指導教員　小田 幹雄 教授
提出年月日　平成 31 年 2 月 20 日

1　はじめに

近年, インターネットの普及により, お客様の問い合わせ手段が増え, クレームや問い合わせ数が増加傾向にある [1]. そのため, クレームや問い合わせに対して素早い対応を行うには, 問い合わせ対応人員の増大など膨大な人件費が必要である. クレームや問い合わせに対して, 人手で行われている知的作業を自動化, 支援することができれば, 多くの企業にとって有益である [2].

本研究では, 共同研究を行っている保険会社の問い合わせ文章データをクレームと非クレームに推定し, 精度の向上を目的としている. 推定には, 単一学習器やアンサンブル学習を用いた学習器など様々な手法を用いた.

2　アンサンブル学習手法

機械学習には, 様々な学習器があるが, いくつかの学習器をまとめて一つの手法としたのがアンサンブル学習である. アンサンブル学習の主な手法にバギングとブースティングがある.

バギング

複数の異なる学習器の予測から, 多数決で推定する方法をバギングという. K 個 (K は奇数) の学習器のアンサンブル学習を考え, 各学習器の誤推定の確率を一律 θ とすると, K 個の学習器のうち,k 個の学習器の誤推定確率 $P(k)$ は

$$P(k) = {}_K C_k \theta^k (1-\theta)^{K-k} \tag{1}$$

となる [3]. 式 (1) より, 過半数の学習器の誤推定確率は単一学習器の誤推定確率より低く, バギングの有効性がわかる.

ブースティング

同じデータ, 学習器を用いて誤った予測をしたデータに重み付けし, 学習する方法をブースティング法という. 用いる学習器を K 個とし, それぞれ $g_i(x)$ と表し, 誤った予測に対しての重みを α_i とすると重み付き平均したアンサンブル学習器の識別関数は,

$$g(x) = \sum_{i=1}^{K} \alpha_i g_i(x) \tag{2}$$

となる.

3　実験

従来, 文書分類問題は,SVM やナイーブベイズといった単一学習器を用いることが一般的であった. しかし, アンサンブル学習を用いれば, 単一学習器で推定するより精度の向上が期待できる.

本実験では, バギングおよびブースティングを用いて, 単一学習器とアンサンブル学習を用いた学習器の精度の比較を行った. まず,SCDV を用いて文章を分散表現ベクトルに変換した. つぎに, 問い合わせデータには非クレーム文章が多く, クレーム文章が少ない傾向があり, 学習器が正確に推定できないため, SMOTEENN を用いてデータの不均衡問題を解決した. バギングには scikit-learn で実装されいるロジスティック回帰, ランダムフォレスト,K 近傍法,SVM, ナイーブベイズの 5 つの学習器を用いており, 各学習器のハイパーパラメータはデフォルトとしている. ブースティングには XGBoost を用いており, ハイパーパラメータを最適化した. また, 学習器の推定精度の評価指標として, 再現率, 適合率,F1 スコアを用いた. 再現率はクレーム文章を正しく推定した確率を表し, 適合率はクレーム文章と推定した文章のうち, 実際にクレーム文章である確率を表す. また,F1 スコアは再現率と適合率の調和平均で求められ, 一般的に F1 スコアが高いほど良い学習器であることを意味している. なお, 本研究では少数のクレーム文章を高精度で推定することが重要であるため, 再現率>90%, 適合率>50% を目標としている. 実験結果を表 1 に示す.

表 1: 各学習器の精度の比較

学習器	再現率 (%)	適合率 (%)	F1 スコア (%)
ロジスティック回帰	75.5245	64.2857	69.4534
ランダムフォレスト	68.1818	71.4286	69.7674
K 近傍法	88.4615	42.8088	57.6967
SVM	52.0979	**85.1429**	64.6421
ナイーブベイズ	**91.2587**	24.0775	38.1022
バギング	74.8252	65.8462	**70.0491**
ブースティング	75.8503	60.4336	67.27

表 1 より,F1 スコアで比較すると, バギングを用いた学習器の方が単一学習器より僅かに精度が向上しており, バギングの有効性が確認できる. また,XGBoost を用いたブースティングが単一学習器よりも F1 スコアが低くなっている. これは,XGBoost のハイパーパラメータの最適化が適切でなかったことが原因として挙げられる. これらの結果を踏まえて, 文章の推定精度を高めるためには, 単一学習器よりもバギングを用いるほうが良く, バギングで用いる学習器はそれぞれ適切にハイパーパラメータを調整する必要があることがわかる.

4　まとめ

本研究では, 単一学習器やアンサンブル学習を用いて, クレームと非クレームの問い合わせ文章の推定を行った. 今後の課題として, アンサンブル学習に用いる学習器の数, 種類を変えることや文章データを特徴エンジニアリングや特徴抽出を用いて加工することで, 精度の向上を図っていくことが挙げられる.

参考文献

[1] OKBIZ.forCommunitySupport, “次世代のサポートチャネル” https://www.okwave.co.jp/business/service/okbiz-cs/option/

[2] 原田 実, 川又 真綱, “クレーム内容の自動分類” 言語処理学会第 14 回年次大会発表論文集 pp.293-294,2008

[3] 上田 修功 “アンサンブル学習” 情報処理学会論文誌 Vol.46 pp11-20,2005