ウェブ上のテキストの書き手の属性推定のための領域適応

鎌田 隆信　　白井 清昭
北陸先端科学技術大学院大学 情報科学研究科
{s1410019,kshirai}@jaist.ac.jp

1 はじめに

近年，ウェブ上のテキストから新しい知識を発見するテキストマイニングや，製品やサービスに関する世間の評判を分析するオピニオンマイニングの研究が盛んに行われている．この際，性別や年齢といったテキストの書き手の属性を推測することができれば，より有用な知識を獲得することができる．例えば，評判情報分析では，女性に人気の商品は何か，高齢者に人気のホテルはどこか，といった分析が可能になる．

本論文では，ウェブ上のテキストの書き手の属性 (性別ならびに年齢) を推定する技術を提案する [3]．教師あり機械学習を用いて書き手の属性を推定する先行研究の多くはブログを対象にしている．これは，ブログではプロフィールの中にブロガーの性別や年齢が書かれていることが多く，訓練データとして必要な正解付きのデータを確保しやすいためと考えられる．一方，本研究では，対象をブログに限定せず，ウェブ上のあらゆるテキストを対象とした汎用的な手法の確立を目指す．

一般に，教師あり機械学習では，訓練データとテストデータの領域が異なると正解率が低下することが知られている．例えば，ブログの訓練データから書き手の属性を推定する分類器を学習した場合，それをブログ記事に適用した場合よりも，一般のウェブページのテキストに適用したときの方が正解率が劣る．訓練データとテストデータの領域が異なるときに正解率を低下させないための技術は領域適応と呼ばれる．本論文では，テキストの書き手の属性を推定するタスクにおける領域適応の手法を提案し，その効果を実験的に評価する．

2 関連研究

書き手の性別や年齢を推定する先行研究の多くは機械学習に基づく．Schler らは，英語ブログを対象に，ブロガーの性別ならびに年齢 (10 代,20 代,30 代) を推定する手法を提案している [5]．機械学習の素性として，機能語の品詞，ブログでよく使われる語，ハイパーリンクといった style-based feature と，自立語やその単語クラスといった content-based feature の 2 つを提案している．性別，年齢判定の正解率はそれぞれ 80%，76%であった．Peersman らは，χ^2 検定によって選択された単語 n-gram や文字 n-gram を素性とし，オランダ語のソーシャルネットワークに投稿されたテキストの著者の年齢・性別を判定している [4]．池田らは日本語ブログの著者の性別を推定する手法を提案している [2]．機械学習の素性として，一人称代名詞，機能語，χ^2 検定で選択された自立語を使用した．男性・女性の二値分類では 88.9%，性別不明クラスを加えた分類では精度 0.93，再現率 0.80 程度の結果が得られた．

先行研究の多くは，ブログやソーシャルネットワークのようにプロフィール欄によって性別や年齢が特定できるデータを対象としている．本研究では，ウェブ上のあらゆるテキストの書き手の性別・年齢を推定することを視野に入れ，属性推定問題における領域適応に焦点を当てる．

3 提案手法

本節では，ウェブ上のテキストが与えられたとき，その書き手の性別ならびに年齢を推定する手法について述べる．性別の推定では「男性」または「女性」を，年齢の推定では「10 代」「20 代」「30 代」「40 代」「50 代」「60 代」の 6 つを分類クラスと定義する．

3.1 書き手の属性の推定

正解付きデータ，すなわち書き手の属性の正しい分類クラスが付与されたテキストの集合を用意する．これを訓練データとし，与えられたテキストに対して性別・年齢を判定する分類器を教師あり機械学習する．学習には LIBLENEAR[1] を用いた．学習の際には，訓練データ中の個々のテキストを素性ベクトルで表現する．素性ベクトルにおける素性の重みは，その素性がテキストに出現すれば 1，そうでない場合は 0 とする．

性別を推定する分類器の機械学習に用いた素性は以下の通りである．

- 一人称の代名詞

[1] https://www.csie.ntu.edu.tw/~cjlin/liblinear/

一人称を表わす代名詞．MeCab の IPA 辞書から品詞が「代名詞」である単語を人手でチェックし，表 1 に示す 36 個の単語を素性とした．

- 付属語列
  テキスト中に連続して出現する付属語の列を素性とする．テキストを MeCab で形態素解析し，品詞によって付属語か否かを判定した．
- 自立語
  テキスト中に含まれる自立語を素性とする．MeCab による形態素解析結果から自立語の品詞を持つ単語を抽出した．また，自立語は素性数が多いため，素性選択を行う．訓練データにおいて，素性と分類クラスの相関を χ^2 値で測り，その上位 N_c 個の自立語のみを素性として利用する．N_c の値は開発データを用いて最適化する．

一方，年齢を推定する分類器の学習に用いた素性は，「付属語列」と「自立語」である．いずれも性別推定のための素性と同じ方法で抽出する．

3.2 書き手の属性推定のための領域適応

一般に，領域適応では，訓練データの領域をソースドメイン，テストデータの領域をターゲットドメインと呼ぶ．ソースとターゲットで領域が異なる場合に分類器の性能が落ちる主な要因は以下の 2 つであると言われている．

- 分類クラスの分布の違い
  ソースドメインとターゲットドメインとで分類クラスの出現頻度の分布が異なるとき，分類の正解率が低下する [6]．例えば，機械学習された分類器はソースドメインにおける最頻出の分類クラスをより多く選択する傾向があるが，それがターゲットドメインではあまり多く出現しないときには分類誤りが多くなる．
- 素性の違い
  ソースドメインとターゲットドメインとで出現する学習素性の分布に違いが見られるときには正解率が低下する [1]．ターゲットドメインの分類に有効な素性があったとしても，それがソースドメインに出現しなければ，分類器に反映されない．また，ソースとターゲットの両方に出現するが，その素性と相関が強い分類クラスが異なる場合には誤分類を生じやすい．例えば，ソースドメインでは男性が書いたテキストによく出現する素性が，

[1] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

あたい, あたし, あたしゃ, あちき, うち, おいら, おのれ, おら, おれ, ぼく, わい, わがはい, わし, わたくし, わたし, わたしゃ, われ, われわれ, ウチ, オイラ, オレ, ボク, 俺, 我々, 我ら, 己, 吾輩, 私, 私しゃ, 私ゃ, 小生, 僕, 僕たち, 僕ら, 僕達, 余

図 1: 一人称の代名詞のリスト

ターゲットドメインでは女性が書いたテキストによく出現する場合が考えられる．

上記の考察を踏まえ，本論文では領域適応のための 2 つの手法を提案する．

3.2.1 分類クラスの出現頻度分布の調整

ソースドメインにおける分類クラスの出現頻度分布がターゲットドメインと同じになるように調整する．これは，ターゲットドメインと比べて出現頻度の大きい分類クラスのデータをソースドメインから削除することで実現する．ターゲットドメインのデータにおける正解の分類クラスがわからないとき，つまり教師なし領域適応の場合には，ターゲットドメインにおける分類クラスの出現頻度を自動的に推定することは難しい．ここでは，何らかの方法でターゲットドメインにおける分類クラスの出現頻度の分布がわかっているものと仮定する．例えば，少数のターゲットドメインのデータに対して，正しい分類クラスを人手で付与する方法が考えられる．この場合，提案手法は半教師あり領域適応となる．

分類クラスが 2 つのとき，ソースドメインにおけるクラス 1，クラス 2 の出現頻度を f_1^s，f_2^s とおく．同様にターゲットドメインにおける出現クラスの出現頻度を f_1^t，f_2^t とおく．ここでの目標は $f_1^s : f_2^s = f_1^t : f_2^t$ が成立するようにソースドメインのデータ数を調整することである．このとき，ソースドメインにおけるクラス 1 の出現頻度の期待値 $f_1'^s$ は

$$f_1'^s = f_2^s \cdot \frac{f_1^t}{f_2^t} \tag{1}$$

で求められる．f_1^s が $f_1'^s$ よりも大きければ，クラス 1 が付与されたデータの中からランダムに $f_1^s - f_1'^s$ 個を選択し，それを訓練データから削除する．もし，f_1^s が $f_1'^s$ よりも小さければ，クラス 2 のデータを削除する．すなわち，クラス 2 のデータ数が $f_2'^s = f_1^s \cdot \frac{f_2^t}{f_1^t}$ となるようにランダムにデータを削除する．

上記の手法を多値分類に拡張したとき，分類クラスの出現頻度の分布を調整するアルゴリズムを Algorithm 1 に示す．Algorithm 1 はソース，ターゲットドメインにおける分類クラスの出現頻度の集合 F_s，F_t を入力

Algorithm 1 分類クラスの出現頻度分布の調整

1: **Input:** $F_s = \{f_1^s \cdots f_N^s\}$, $F_t = \{f_1^t \cdots f_N^t\}$
2: **Output:** $F'_s = \{f_1'^s \cdots f_N'^s\}$
3: **for all** i such that $1 \leq i \leq N$ in descending order of f_i^s **do**
4: $\quad f_i'^s \leftarrow f_i^s$
5: $\quad$ **for all** j such that $1 \leq j \leq N$ **and** $j \neq i$ **do**
6: $\quad\quad f_j'^s \leftarrow f_i^s \cdot \frac{f_j^t}{f_i^t}$
7: $\quad\quad$ **if** $f_j'^s > f_j^s$ **then**
8: $\quad\quad\quad$ unset all values in F'_s ; goto LINE 3
9: $\quad\quad$ **end if**
10: $\quad$ **end for**
11: $\quad$ **return** F'_s
12: **end for**

とし，調整後のソースドメインにおける分類クラスの出現頻度の集合 F'_s を返す．まず，基準となる分類クラス i をひとつ選択する．基準クラスはソースドメインにおける出現頻度 f_i^s の大きい順に選択する (3 行目)．基準クラス i の調整後の出現頻度は元の出現頻度と同じとする (4 行目)．基準クラス以外の分類クラス j に対し，式 (1) と同様の式で出現頻度の期待値 $f_j'^s$ を求める (6 行目)．これが実際の出現頻度 f_j^s よりも大きいとき，データを削除することによってクラス j の出現頻度を $f_j'^s$ に合わせることは不可能なので，これまでの計算結果を全て破棄し，基準クラスを変更する (7〜9 行目)．全ての分類クラスについて $f_j'^s \leq f_j^s$ を満たすような $f_s'^j$ を決めることができたら，F'_s を返す．その後，ソースドメインの訓練データにおける各クラスの出現頻度が F'_s と一致するようにデータを削除する．削除するデータはランダムに選択する．

3.2.2 有効性の低い素性の削除

ターゲットドメインのデータの分類に有効ではない素性を削除する．素性 f に対し，ソースドメイン，ターゲットドメインにおいて f と最もよく共起する分類クラスをそれぞれ $C_s(f)$，$C_t(f)$ とおく [2]．もし，$C_s(f)$ と $C_t(f)$ が異なるとき，その素性はターゲットドメインにおける書き手の属性の推定を誤る要因になりうる．ここでは，$C_s(f)$ と $C_t(f)$ が異なるような素性を有効性の低い素性と定義し，これを削除する．具体的には，まず各素性に対して $C_s(f)$, $C_t(f)$ を求める．$C_t(f)$ を求める際，ターゲットドメインのデータの分類クラスは，ソースドメインを訓練データとして学習した分類器を用いて自動推定する．そして，以下の条件を満たす素性 f を削除する．

1. $C_s(f) \neq C_t(f)$
2. f のソースドメインにおける出現頻度が 5 以上である．
3. 上記 1,2 の条件を満たす素性を χ^2 値の大きい順に並べたとき，f はその上位 N_f 件に含まれる．

4 評価実験

4.1 実験データ

ソースドメインのデータとして，Yahoo!ブログ [3] のブログ記事を用いた．Yahoo!ブログのカテゴリの中から 190 件を選択し，カテゴリ毎に最新のブログ記事を 1000 件ダウンロードした．プロフィール欄に性別や年齢が明記されている場合，その性別・年齢を正解の分類クラスとみなし，明記されていないブログ記事は除外した．ターゲットドメインのデータとして，楽天トラベル [4] に投稿されたホテルのレビュー文を用いた．人気上位 5 件のホテルに対して投稿されたレビュー文を取得した．ブログ記事と同様に，投稿欄に明記されている性別や年齢から正解の分類クラスを付与した．ソースドメイン，ターゲットドメインのデータ数を表 1 に示す．

表 1: 実験データ

	性別	年齢
ブログ (ソース)	120,597	17,547
レビュー (ターゲット)	6,985	6,951

4.2 実験結果

提案手法の評価のために，以下の 6 つの設定で性別ならびに年齢を判定した．ただし，最初の 3 つの設定 (B-B，R-R，B-R) では領域適応の技術は用いない．それ以外の設定では，ソースをブログ，ターゲットをレビューとし，領域適応の技術を用いてターゲットデータの性別・年齢の推定を試みる．

B-B ブログのみを使用する．ブログデータを 8:1:1 に分け，それぞれを訓練，開発，テストデータとした．開発データは N_c の最適化のために用いる．

R-R レビューのみを使用する．楽天トラベルのデータを 9:1 に分割し，それぞれを訓練・テストデータとした．

[2] 一般には，素性 f と分類クラスの共起頻度が最大かつ等しくなるような分類クラスは複数存在するので，$C_s(f)$, $C_t(f)$ はともに分類クラスの集合である．

[3] http://blogs.yahoo.co.jp/

[4] http://travel.rakuten.co.jp/

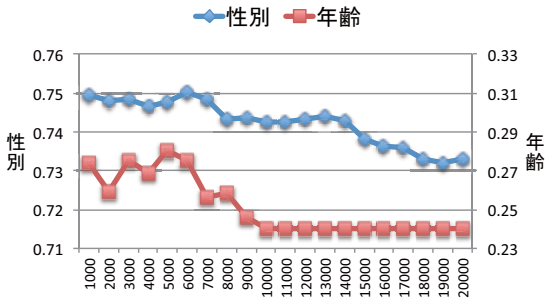

図 2: N_c の最適化

B-R ブログを訓練データ，レビューをテストデータとして使用する．

B-R-C100 3.2.1 で述べた分類クラスの出現頻度分布を調整する手法で領域適応する．ターゲットドメインから 100 個のデータをランダムに選択し，それを用いてターゲットドメインの分類クラスの出現頻度の分布を得る．この試行を 5 回繰り返したときの正解率の平均を示す．

B-R-Cgold B-R-C100 と同様だが，全てのターゲットドメインのデータを用いて分類クラスの出現頻度の分布を得る．

B-R-FS 3.2.2 で述べた有効性の低い素性を削除する手法で領域適応する．

まず，自立語の素性選択のためのパラメタ N_c の最適化を行った．最適化はブログのデータを用いて行う．N_c の値を変動させたときの開発データにおける正解率の変化を図 2 に示す．この結果から，性別では $N_c = 6000$，年齢では $N_c = 5000$ と設定した．

次に，性別および年齢推定の実験結果を表 2 に示す．B-B および R-R はソースとターゲットのドメインが同じ場合，B-R はソートとターゲットのドメインは異なるが領域適応をしない場合である．一般には前者の方が正解率が高くなるが，今回の実験では B-B における性別の正解率は B-R よりも低くなっている．原因は不明だが，ブログでの性別判定よりもレビューでの性別判定の方が易しい可能性がある．

分類クラスの出現頻度分布を調整する手法 (B-R-C100) は，B-R と比べて，性別推定では 0.7%，年齢では 3.3%，正解率が向上している．また，ターゲットドメインの出現頻度を求める際にターゲットドメインの全データの正解クラスを利用したとき (B-R-Cgold) と比べて，B-R-C100 の正解率はそれほど変わらない．このことから，ターゲットドメインの分類クラスの分布を調べるためには 100 個程度のサンプルでも十分ということがわかる．

表 2: 書き手の属性推定の正解率

	性別	年齢
B-B	0.766	0.239
R-R	0.686	0.350
B-R	0.638	0.281
B-R-C100	0.645	0.314
B-R-Cgold	0.637	0.315
B-R-FS (N_f=100)	0.627	0.300
B-R-FS (N_f=200)	0.627	0.305
B-R-FS (N_f=217)	0.628	–
B-R-FS (N_f=500)	0.628	0.302

有効性の低い素性を削除する手法 (B-R-FS) については，削除する素性の数 N_f を変えて実験を行った．その結果，性別推定の場合では $N_f = 217$[5]，年齢推定の場合では $N_f = 200$ のときに正解率が最も高くなった．性別推定では B-R よりも正解率が悪くなったが，年齢推定では正解率が 2.4%向上した．

5 おわりに

本論文では，単純な領域適応の手法を提案し，それを書き手の性別・年齢を推定するタスクに適用したときの実験結果を報告した．提案手法と既存の領域適応の手法との比較や，提案手法を既存手法と組み合わせて用いる方法を探究することが今後の課題となる．また，今回の実験では，年齢推定の正解率は全般的に低かった．年齢推定のための有効な素性を発見し，正解率を向上させることにも取り組む必要がある．

参考文献

[1] Hal Daumé III. Frustratingly easy domain adaptation. In *Proceedings of ACL*, pp. 256–263, 2007.

[2] 池田大介, 南野朋之, 奥村学. blog の著者の性別推定. 言語処理学会第 12 回年次大会, pp. 356–359, 2006.

[3] 鎌田隆信. 評判分析のための著者の性別及び年齢の推定. Master's thesis, 北陸先端科学技術大学院大学, 3 2016.

[4] Claudia Peersman, Walter Daelemans, and Leona Van Vaerenbergh. Predicting age and gender in online social networks. In *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents*, pp. 37–44, 2011.

[5] Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James Pennebaker. Effects of age and gender on blogging. In *Proceedings of AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pp. 199–205, 2006.

[6] 新納浩幸, 佐々木稔. 共変量シフト下の学習による語義曖昧性解消の教師なし領域適応. 自然言語処理, Vol. 21, No. 5, pp. 1011–1035, 2014.

[5] これは，$C_s(f) \neq C_t(f)$ となる全素性を削除したときである．