社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－ＦＩ－ 89 (9)
2008－ＮＬ－183 (9)
2008／1／22

blog分類のための半教師有り学習

池田 大介† 高村 大也‡ 奥村 学‡

概 要

blog著者の属性推定など, 教師有り学習を用いblogを分類する研究がなされている. ラベルの無いblogであれば容易に収集が可能であるが, 正解ラベル付きのblogは一般に高価である. そこで, 本研究では半教師有り学習によるblog分類手法を提案する. blog中の各エントリはスタイルや内容が共通している. 本研究ではこれに着目し, 各エントリがどのblogに属していたか, という補助問題を解くことにより, blogのスタイルやコンテンツと言った各blogに固有の特徴をモデル化する. この情報を利用することで, 目的の分類問題の精度を向上させることができる. 本手法を用いた, いくつかの分類タスクでの実験結果についても報告する.

Semi-supervised Learning for Blog Classification

Daisuke IKEDA† Hiroya TAKAMURA‡ Manabu OKUMURA‡

Abstract

Classifying blogs, e.g. identifying bloggers' gender or age, is one of the most interesting problems in blog analysis today. Although it is usually solved by applying supervised learning techniques, it is not always easy to collect labeled blogs enough to train an accurate classifier. To the contrary, we can collect a huge amount of blogs that have no labels. In this paper, therefore, we propose a semi-supervised learning method for blog classification in order to incorporate unlabeled data into supervised learning. We assume that the entries from the same blog have the same characteristics. With this assumption, our method captures the characteristics of each blog, such as writing styles, and uses it to improve classification accuracy.

1 背景

近年, blogの急速な普及に伴いその情報源としての期待が高まってきており, blogを対象とした自然言語処理に関する研究も数多くなされている. blogのトピック分類やスパムフィルタリングといった従来の自然言語処理技術をblogに適用したタスク ([7],[5]) や, blog著者の性別や年齢, 居住地といった属性を推定する ([1], [2], [4], [6]), といったblog固有のタスク等がその例である. これらのタスクは, "blogを分類する" 問題であると見なすことができる.

blogはエントリと呼ばれる文書の集合であることから, blogの分類問題は文書分類の問題として考えられることが多い. 文書分類は古くより教師有り学習の枠組みが良い結果を残しており, blogの分類でも同様のアプローチによる手法が主流である. しかし, 教師有り学習で十分な精度を得るには, 多量の正解ラベル付きの事例を用いた学習が必要である. つまり, 分類タスクごとに十分な量の正解ラベル付きblogを用意しなければならない. こういったデータは必ずしも入手が容易ではなく, コストが高くなってしまうことが多い. 一方で, blogはWebから容易に収集が可能という性質がある. 現在Web上には数十万, 数百万というblogが存在していることと合わせると, 正解ラベルの無いblogであれば安価に, かつ大量に入手可能であると言える. つまりblogの分類は, 少量しか無いラベル付き事例を多量のラベル無し事例で補う, 半教師有り学習の枠組みに適していると言える.

blogの分類では, 従来自然言語処理が対象としてきた新聞記事の分類などと異なり, ライティングスタイルや記述されるコンテンツの傾向といった各blogに固有

†東京工業大学大学院 知能システム科学専攻
Department of Computational Intelligence and System Science, Tokyo Institute of Technology
ikeda@lr.pi.titech.ac.jp
‡東京工業大学 精密工学研究所
Precision and Intelligence Laboratory, Tokyo Institute of Technology
{takamura,oku}@pi.titech.ac.jp

な特徴が重要な意味をもつタスクが存在する．年齢や性別の推定がその一例である．

本研究では，ライティングスタイルのような各 blog の特徴を考慮した，半教師有り学習による blog 分類手法を提案する．本手法では大量の正解ラベルの無い blog から各 blog の特徴を捉え，これを特徴量として目的とする分類問題の学習に利用する．本手法を用いた blog 著者の性別推定，年齢推定の結果についても報告する．

1.1 blog の個性

blog のライティングスタイルやトピックの傾向は，その blog のカテゴリや著者の属性といった，blog の分類の基準となるような情報と密接な関係がある．例えば，プログラマの blog には技術的な話題が多く出現するかもしれない．政治に関する blog であれば砕けた表現は好まれず，新聞記事に近いライティングスタイルで記述されるだろう．一部の顔文字や記号は特定の年代の人間を中心に用いられている．

こういったライティングスタイルのような特徴は各 blog に固有な性質であるため，本稿ではこれを blog の個性と呼ぶことにする．blog の個性は分類において有用な情報であるにも関わらず，従来の blog 分類手法では考慮されることが無かった．本研究ではラベル無し事例からこれを捉える手法を提案する．

1.2 半教師有り学習

半教師有り学習とは，正解ラベルのある事例とそうでない事例の両方を用いた学習を意味するが，厳密な定義は諸説ある．EM アルゴリズムのように学習の際にラベル無し事例を直接用いる手法のみを指すという説もある．

本稿で指す半教師有り学習はそのような必要条件を考えておらず，いわば広義の半教師有り学習を意味する．すなわち，ラベル有り事例とラベル無し事例を用いた学習法であれば，特にその利用法に制限は設けない．

2 関連研究

blog はリアルタイム性に優れ，個人の意見等が多く記述されるという特徴がある．これを利用して，blog からある商品が世間で好評であるか否か，といった評判情報や，現在流行している物は何か，といった情報を獲得することを目的とした研究が行われている [15], [14]．これに伴い，blog を分類するというタスクについても研究が行われている．スパム blog を排除するための分類 [7] や，それぞれの blog がどのような属性を持つ人によって記述されたかによる分類 ([1], [2], [4], [6]) 等である．これによって分析対象とすべきでない blog を取り除くことや，評判や流行の分析を年齢や性別ごとに行うことが可能となる．このような技術を用いたアプリケーションは少数ながらすでにサービスとして運用されているものもあり[1,2]，blog の分析技術は今後より普及していくと思われる．

しかし，先にも述べた通り blog を分類するというタスクはこれまで教師有り学習の問題として解かれてきた．すなわち，blog をそれに含まれる全てのエントリをつなげた長い文書と見なし，文書分類手法を適用する．教師有り学習は古くより良い結果を残している確立された手法ではあるが，Web や blog の豊富な言語資源を生かすには半教師有り学習の技術が必須である．

半教師有り学習による文書分類手法は古くから研究がなされている．EM アルゴリズム [16] や共訓練 [17] を用いた手法は広く知られている．blog の分類もこれらの手法を用いる事で半教師有り学習に拡張可能であるが，これらはあくまで文書分類のためのアルゴリズムである．本稿で提案する手法は blog の特徴を利用したものであり，従来法と比べ汎用性は劣るものの，より目的とするタスクに特化した手法であると言える．

提案手法は Ando ら [8] の提案した，Alternating Structure Optimization(ASO) の応用であると見なすことができる．ASO は半教師有り学習やマルチタスク学習に適用可能な機械学習のフレームワークであり，文書分類 [8] や語義曖昧性解消 [10]，POS タギング等のチャンキング [9]，意味役割付与 [11]，ドメイン適応 [12][13] といった様々な分野に応用した研究が報告されている．ASO の詳細と本研究との関連については，4 節で詳しく述べる．

3 用語の定義

本稿で提案する手法は，新聞記事や通常の Web ページにはない blog の独自の特徴を利用する．混乱を避けるため，まず本節で blog に関する用語の紹介と定義を行う．

- **blog**: 特定の人が文書を投稿可能で，それらを時系列順に並べて閲覧可能な日記的な Web サイト．blog サイトとも呼ばれる．本稿では，一つの blog を一つの事例として分類する問題を考える．
- エントリ: blog に対する一回の投稿．記事，パーマリンクなどと呼ばれる事もある．blog はエントリという短い文書の連なったものと言える．

[1] blogWatcher http://blogwatcher.pi.titech.ac.jp/
[2] blogeye http://blogeye.jp/

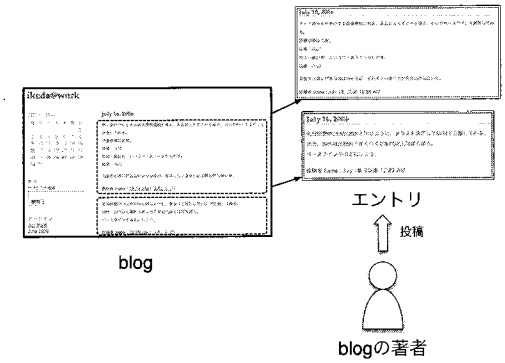

図 1: blog とエントリ, 著者の関係

- **blog の著者**: blog を管理, 運営している人. エントリの投稿やレイアウトの編集を行う. 多くの blog は一人の著者によって運営されている.

図 1 のように, ある人が blog を開設し, 著者としてそこへエントリを日々投稿する, という流れが一般的である.

4 提案手法

4.1 副分類器

先に述べた通り, blog の個性は分類において有用な情報になりうるが, 定量的な表現が難しく数値化が困難であるため従来手法では積極的に考慮されてこなかった. 仮にこれが数値化できていれば, その blog の一種の特徴量として機械学習による手法に取り込む事で blog の個性を考慮した分類が可能である. これまでの研究では blog の個性を間接的に表現するため, それと密接な関係があると考えられる各助詞の頻度や用いられる一人称代名詞等が特徴量として用いられた [1]. 本研究では, blog の個性を他の blog との類似性を基に表現する. 例えば, blog A は blog B には似ていないが, blog C には似ている, といった表現方法である. ある blog が他のどの blog に似ており, どの blog に似ていないかという情報はその blog の個性をある側面から数値化した情報であると言える. 同じ blog に似ている blog 同士もまた似た blog であるだろう.

この表現を実現するためには, 入力された blog がどの blog にどの程度似ているかを判定する仕組みが必要である. しかし, 2 つの blog のみからそれらが似ているか否かという絶対的な判定は難しい. 似ているか否

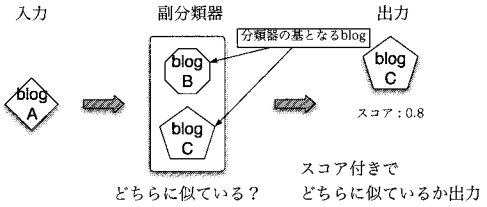

図 2: 副分類器の例

かという判断は他にどういったサンプルが存在するかという情報無しにはできない. 例えば, サッカーに関する blog と野球に関する blog は blog 全体から見れば似ていると言えるだろうが, スポーツに関する blog の間では大きく異なる blog と見るべきだろう. そこで本研究では, 絶対的に 2 つの blog が似ているか否かを測るのではなく, blog A は blog B よりも blog C に似ている, というような相対的な観点で類似性を捉える. これは, 入力された blog が基となる 2 つの blog のうちどちらに似ているかを出力する分類器を用いる事で実現可能である.

具体的には, $\boldsymbol{x}$ を入力される blog のベクトル表現として, 以下のような線形識別モデルを考える:

$$y = \boldsymbol{u} \cdot \boldsymbol{x}. \tag{1}$$

y はどちらの blog にどの程度近いかを表すスコアである. 正負でどちらに近いかを, 絶対値でどの程度近いかを表現する. $\boldsymbol{u}$ はモデルのパラメータであり, 基となる 2 つの blog から学習により求める. 詳細な学習法については後述する.

本研究ではこの分類器を, 本来求めるべき分類器の為に構築するものであることから, "副分類器" と呼ぶ. 図 2 に副分類器のイメージを示す.

副分類器はライティングスタイルを始めとする blog の個性だけを考慮するため, 最終的な分類タスクには非依存である. すなわち, この副分類器はラベル無し事例を基に学習することが可能であり, ラベル無し事例は収集が容易であることから, 大量の分類器を構築することができる. 言い換えれば, 本研究では各 blog の個性を捉えるため他の blog との類似性を考えるが, その類似性を測るための blog は目的とするタスクに関するラベル付きの blog である必要はない. 大量のラベル無し blog との類似性を考えることで, より正確に blog の個性を捉える事が可能になると思われる.

4.2 副分類器の学習法

次に, 副分類器をどのように学習するかについて説明する. どの blog がどちらに似ているという正解ラベルは存在しないため, 通常の教師有り学習手法は適用できない. 唯一, 基となる 2 つの blog がそれぞれ自分自身に似ているということは解る (例えば blog A と blog B を基にするのであれば, blog A は blog A 自身に, blog B は blog B 自身に似ている事だけは自明である) が, 2 事例だけから満足な学習は難しい.

そこで本研究では, **blog は複数のエントリで構成される**という点に着目する. 基となる 2 つの blog 中の各エントリを事例と見なし, 一方の blog のエントリを正例, もう一方の blog のエントリを負例として式 (1) のモデルの学習を行う. この設定であれば, 2 つの blog からであっても, 十分な数の訓練事例を得る事ができる. また, どのエントリがどの blog 中に記述されたものであるかは, 通常 blog の収集時に判明している. つまり, この設定であれば正解ラベルを一切必要とせずに, 収集された任意の 2 つの blog から分類器を学習できる.

これによって得られる分類器は, 入力されたエントリが, 基となる二つの blog のどちらに属していたか, を分類するモデルと見ることができる. ここで, **同じ blog 中のエントリは, ライティングスタイル等, blog の個性が共通している**という仮定を置く. 通常, blog は一つ一つ別の人間によって運営されており, そのライティングスタイルを始めとする blog の個性は様々である. 一方, blog は多くの場合同じ人間が継続的に書いた複数のエントリの集合である, ということに着目すると, それらには同じ個性が現れていると考えることができる.

つまり, この分類器は 2 つの blog の個性の差異を学習することが期待できる. また, この分類器で他の blog を分類することにより, それが基となった 2 つの blog のどちらに近いかを測ることが可能である.

4.3 副分類器の学習の必然性

ここまでで副分類器の学習法について述べたが, 相対的な類似度を求める, という目的だけを考えれば学習を行う必然性は無い. 例えば, 古くから情報検索等で用いられているコサイン類似度や内積によって類似度を計算し, それを相対的な値へ変換する事でも相対的な類似度は計算可能である.

しかし, 提案手法である学習を行う事により, 基となっている 2 つの blog の差異を捉える事が可能となる. 例えば, サッカーに関する blog と野球に関する blog を基にした副分類器を考えよう. これら 2 つの blog は共通する部分が多いため, 学習を用いないコサイン類似度や内積ではどのような入力に対しても, どちらとも判断がつかないという出力を得るだろう. しかし, 学習を行えば, 2 つの blog を分離する際に有効な素性に大きな重みをつける事ができる. 先の例で bag-of-words を素性として用いれば, “バット”, “オフサイド” といった一方の blog ではよく出現するような素性に対して大きな重みが付く事が期待できる. これによって, 入力された blog がどちらに近いかをより明確にすることができる.

4.4 半教師有り学習による blog 分類

提案手法である半教師有り学習による blog 分類手法について説明する. 副分類器の出力は, 入力された blog が基となった 2 つの blog のどちらに近いかを表す. これはある側面から blog の個性を数値化したものと考えられる. 大量のラベル無し blog を用い, 大量の分類器を作ることで, 様々な側面から評価した blog の個性を得る事ができる. つまり, この出力を並べたベクトルは, 入力された blog の個性を捉えた特徴ベクトルになっていると言える.

こうして作成された blog の特徴ベクトルを, 本来解くべき分類タスクのための素性として加える. つまり分類の入力ベクトルは, 副分類器の数だけ次元の増えたベクトルとなる. これを用いて学習を行うことで, ライティングスタイルを考慮した学習が可能である. またこの手法は, ライティングスタイルのベクトル表現のためにラベル無し事例の情報を積極的に取り入れた半教師有り学習でもある.

この学習法は以下の様に書く事もできる. 今, 副分類器のモデルが K 個構築できているとする (それぞれ, $\boldsymbol{u}_0, ..., \boldsymbol{u}_{K-1}$ とする). これらを並べた行列を $U \equiv [\boldsymbol{u}_0, ..., \boldsymbol{u}_{K-1}]$ とすると, 各モデルの出力を並べたベクトルは $U^T\boldsymbol{x}$ で表せる. これを元の入力ベクトル $\boldsymbol{x}$ と並べ, $\left\{\left[\boldsymbol{x}_i^L, U^T\boldsymbol{x}_i^L\right], y_i\right\} (i = 1...N)$ を訓練事例として学習する. モデルは副分類器同様, 以下の線形識別モデルを用いた:

$$y_i = \boldsymbol{w} \cdot \left\{\left[\boldsymbol{x}_i^L, U^T\boldsymbol{x}_i^L\right], y_i\right\} (i = 1...N) \qquad (2)$$

分類の際には, 訓練時同様副分類器の出力をそれぞれ計算し, それを素性として追加したベクトルを分類すればよいが, この方法では副分類器の数だけ分類する必要があるため, 実行速度は副分類器の数に対し線形に増える. しかし, U は問題によらず固定であるため, $\boldsymbol{w}U^T$ をあらかじめ計算しておく事により, 分類時間を副分類器に対しては定数時間にすることが可能である.

多くの学習アルゴリズムは各素性や特徴量に対し重みを付与し, それを用いてスコアを計算する事で分類を

入力：ラベル有り blog$\{(\boldsymbol{x}_i^L, y_i)\}(i=1,...,N)$,
　　ラベル無し blog$\{\boldsymbol{x}_i^U\}(i=1,...,M)$.
　　各 blog $\boldsymbol{x}$ には，エントリ $\{\boldsymbol{e}\} \subset \boldsymbol{x}$ が含まれる．
パラメータ：副分類器の数 K
出力：分類器のパラメータ $\boldsymbol{w}$
1: FOR $l=0$ TO $K-1$
2: 　相違なるラベル無し blog を
　　ランダムに 2 つ選択
　　（それぞれ $\boldsymbol{x}_s^U$, $\boldsymbol{x}_t^U$ とする）．
3: 　$\{\boldsymbol{e}_l^s\} \subset \boldsymbol{x}_s^U$ と $\{\boldsymbol{e}_l^t\} \subset \boldsymbol{x}_t^U$ を分類する
　　モデル（＝副分類器）$\boldsymbol{u}_l$ を学習．
4: NEXT
5: $U \equiv [\boldsymbol{u}_0, \boldsymbol{u}_1, ..., \boldsymbol{u}_{K-1}]$
6: $\left\{\left[\boldsymbol{x}_i^L, U^T\boldsymbol{x}_i^L\right], y_i\right\}(i=1...N)$ を
　　訓練事例とし，$\boldsymbol{w}$ を学習．

図 3: 提案手法のアルゴリズム

行う．提案手法によって追加された特徴量に対し与えられる重みは，その特徴量を与えた副分類器と解きたい分類タスクとの相関を表していると言える．例えば，解きたい分類タスクが blog 著者の性別の分類であった場合，著者の性別の異なる 2 つの blog を基に作られた副分類器に対して大きな重みが与えられると思われる．

本節で提案した半教師有り学習のアルゴリズムを図 3 にまとめる．

5 Alternating Structure Optimization としての提案手法

先にも述べた通り，提案手法は Ando ら [8] による ASO の応用の一つであると見なすことができる．本節では ASO のアルゴリズムについて説明し，ASO の応用として見た提案手法について述べる．

5.1 ASO の概要

先にも述べた通り，ASO はマルチタスク学習や半教師有り学習のためのフレームワークであり，文書分類や POS タギングなど様々な応用が報告されている．ASO では，本来解くタスクとは別に auxiliary problem と呼ばれる問題を用いることで半教師有り学習を可能にしている．どのような auxiliary problem を用いるかが学習の結果に大きく影響する．auxiliary problem の例として，Ando ら [8] は以下のようなものを提案している．

- 頻出語の推定
  文書分類を解く際の auxiliary problem．共学習のように単語空間を 2 つにわけ，一方の単語集合のみから，もう一方の集合のうち入力された文書で最も出現頻度の高い単語を推定する．
- 系列中の語の推定
  POS タギングなど，系列ラベリングの際に用いる事が出来る auxiliary problem．事例中の単語を隠し，前後の系列からその単語を推定する．

これらを，大量の 2 値分類問題として解く．頻出語の推定であれば単語の種類数と同数の分類問題を作る事ができる．また，これらの問題は目的とするタスクの正解ラベルを用いずに学習が可能であるため，ラベル無し事例を用いて学習する事が可能である．この大量の問題をマルチタスク学習の要領で同時に解く事により，素性間の関係といった通常の学習では考慮できない情報を発見し，学習に取り込むのが ASO アルゴリズムである．具体的には以下の手順で学習を行うことができる．

1. auxiliary problem による各 2 値分類問題をラベル無し事例から学習し，$\boldsymbol{u}_i$ を得る．
2. $U \equiv [\boldsymbol{u}_0, \boldsymbol{u}_1, ...]$ とする．
3. 特異値分解によって $U = V_1 D V_2^T$ と分解し，Θ を V_1 の最初の h 列とする．
4. 訓練事例 $\{\boldsymbol{x}_i, y_i\}(i=1...N)$ を $\left\{\left[\boldsymbol{x}_i, \Theta^T\boldsymbol{x}_i\right], y_i\right\}(i=1...N)$ と変換し，学習する．

なお，特異値分解による変換と圧縮は理論的な正当化がなされているが，予備実験の結果大きな効果が得られなかったため，本研究では用いていない．

5.2 提案手法との関係

本稿で提案した手法は，ASO を blog に適用し，blog の分類というタスクに適した副分類器という auxiliary problem を用いた手法と言える．

Ando らは auxiliary problem には 2 つの条件が必要であると述べている．すなわち，

- Automatic Labeling
  ラベル無し事例から自動的にラベル有り事例を生成可能
- Relevancy
  本来解くタスクとある程度関連がある．

の 2 つである．なお，2 つのタスクが関連がある，とは，一方を解く際にもう一方の情報に利用価値がある，と言い換える事もできる．

本研究で提案した副分類器もまた，この 2 つの条件を満たしており，auxiliary problem と見なす事が出来る．

- Automatic Labeling
  どのエントリはどのblogに属しているか，という情報は収集時に判明しており，自動的にラベル有り事例を生成可能．
- Relevancy
  副分類器から得られるblogの個性に関する情報は目的のタスクを解く上で有用と思われる．ランダムに2つ選んだblogから副分類器を構成する際，選ばれた2つのblogが目的のタスクと合致していた場合，高い関連性が得られる．例えば性別推定の問題を解く際に，男性のblogと女性のblogから作られた副分類器は有益な情報を含んでいるだろう．

5.3 他のタスクへの応用

従来のauxiliary problemには素性の一部を他の素性から推定するという素性間の関係を学習するものが多い．我々の副分類器はラベル無し事例の間の関係を学習しているという点で従来のauxiliary problemと異なる．複数のエントリで構成される，同じblog中のエントリはblogの個性が共通している，といったblogの特徴や仮定が，これを可能にしている．同様のauxiliary problemを他のタスクに適用するためには，次のような条件が必要である．

- 分割可能である．
  一つの事例を複数の事例に分割可能．blogの場合エントリ単位に分割可能．
- 分割された事例がそれぞれ共通の性質を持つ．
  共通する性質がなければ，分割された事例が元々どの事例から得られたものであるか，という分類問題を解く事ができない．blogでは全てのエントリで個性が共通という仮定を置いた．
- その性質が目的の問題と関連がある．
  auxiliary problemの条件であるrelevancyを満たすために必要な条件．blogの場合は個性が分類問題に有効という点で満たされていた．

6 評価実験

提案手法の有効性を確認するため，評価実験を行った．タスクとしてblog著者の性別推定，年齢推定についてそれぞれ実験した．性別推定は，各blogをその著者の性別が男性であるか，女性であるか，の2クラスに分類する問題，年齢推定は著者の年齢を10代から50代までの5つのクラスに分類する問題である．これらはblog著者の属性推定のうちで基本的な問題であることから，これまでにも研究がなされている[1][2][6]．

6.1 データセット

2つの実験でほぼ同様のデータセットを用いた．正解ラベル付きblogとして，Yahoo! blog[3]より，プロフィール欄に性別，年齢の記述のあるblogを収集し，これを用いた．ラベル無し事例は，Livedoor blog[4]より収集した．これらの事例から，無作為に組を作り，副分類器を学習した．

なお，blogサービスによってユーザ層やコミュニティー等が違うことが考えられるため，同じblogサービスからラベル付き，ラベル無し両方のblogを収集した方が高い精度が得られる可能性はある．また，手法の厳密な分析のためにはこのようなデータの差異は極力避けた方が望ましいだろう．しかし現実には各タスク毎にラベル無しblogを収集し直すというのは現実的ではなく，特定のラベル無しblogしか用いる事ができないのではラベル無し事例は容易に，大量に用いる事ができるという利点が損なわれる．こういった理由から本稿では収集元の異なるblogを用いて半教師有り学習を行った．

6.2 実験設定

副分類器の学習にはパーセプトロンベースのオンライン学習アルゴリズム[3]を用いた．大量の副分類器を学習する必要があるが，同アルゴリズムは高い精度を保ちながら高速であると知られているためである．目的とするタスクの学習には頑健性を重視し，サポートベクターマシン(SVM)を用いた．また，素性としては副分類器，性別推定，年齢推定全ての分類器でbag-of-wordsを用いた．つまり，提案手法ではbag-of-wordsベクトルに，副分類器の出力である実数値ベクトルを繋げたベクトルを入力ベクトルとする．また，全ての実験で線形カーネルを用いた．

評価指標としては正解率を用いた．

手法として，以下の3種を比較する．

- ベースライン: ラベル無し事例を用いない，教師有り学習による手法．
- **提案1000**: 提案手法．副分類器を1,000個用いる．
- **提案10000**: 提案手法．副分類器を10,000個用いる．

[3] http://blogs.yahoo.co.jp/
[4] http://blog.livedoor.com/

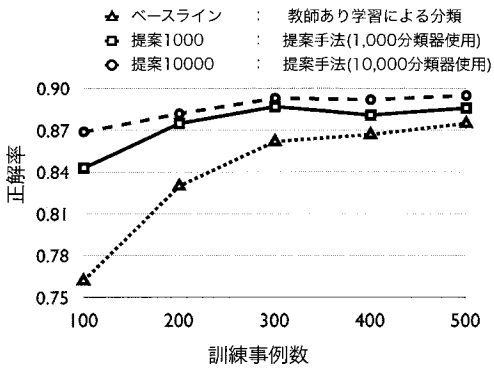

図 4: 性別推定の結果

表 1: 年齢推定の結果

	one-versus-rest	pairwise
ベースライン	0.629	0.627
提案 1000	0.636	**0.642**
提案 10000	0.616	0.635

6.3 blog 著者の性別推定

まず, 性別推定の結果について報告する. 2112 例をテスト事例とし, 訓練事例数を 100 から 500 まで段階的に変化させ, 評価した.

図 4 がその結果である. 訓練事例数によらず提案 10000, 提案手法 1000 がベースラインより良い結果を得ており, 提案手法の有効性を確認できた. この改善は, 訓練事例数が少ない場合に特に顕著であり, 提案手法は少量のラベル付き事例しか入手できないようなタスクであっても効果的に働くと言える. また, 提案 10000 が提案 1000 より高い正解率を得ていることから, より多くのラベル無し事例の情報を用いることで, より良い結果が得られ, さらに多くのラベル無し事例を用いることでさらなる性能の向上が期待できる.

6.4 blog 著者の年齢推定

次に, 年齢推定の結果について報告する. 訓練事例として 2000 事例, 評価事例として 1314 事例を用いた. また, 年齢推定は 10 代から 50 代までの 5 クラス分類であるため, SVM を one-versus-rest と pairwise の 2 種類の方法で拡張し, 多クラス分類器として用いた. 結果は表 1 に示す. 提案 1000 は, one-versus-rest, pairwise

表 2: クラス毎の年齢推定の結果

	ベースライン	提案 1000	提案 10000
10 - 20	0.863	0.866	0.864
10 - 30	0.893	0.919	0.917
10 - 40	0.945	0.951	0.954
10 - 50	0.946	0.943	0.946
20 - 30	0.751	0.769	0.736
20 - 40	0.908	0.912	0.908
20 - 50	0.957	0.960	0.962
30 - 40	0.790	0.792	0.747
30 - 50	0.871	0.893	0.906
40 - 50	0.703	0.703	0.734

両方でわずかだがベースラインを上回っている. しかし, 提案 10000 は pairwise ではベースラインを超えているが提案 1000 を下回り, one-versus-rest ではベースラインをも下回る結果となった. ベースラインは one-versus-rest と pairwise の間の差は小さいが, 提案手法はどちらも pairwise が良い結果を得た. これは提案手法で用いている副分類器が 2 つの blog を比べるという pairwise に近い設定であるためと思われる.

提案手法があまり良い結果を得られなかった原因を調べるため, 各クラスでの正解率について考察する. 表 2 は pairwise における, 年齢推定のクラス毎の結果である. 一番左の列はどのクラスとどのクラスを分類しているかを表す. pairwise を用いているので, 2 クラスずつ分類している. 例えば “10 - 30” の行は, 10 代か 30 代かを分類する分類器に対する結果である. なお, 正解率で評価しており, 分類器が対象としているクラスの事例のみで評価している. 例えば, “10 - 30” の行の結果は, 10 代によって書かれた blog と 30 代によって書かれた blog のみをテスト事例として評価している.

これを見ると, 20 代-30 代, 30 代-40 代といったクラスで提案 10000 が極端に悪い結果となっていることが解る. この理由として, これらのクラスはライティングスタイル等の個性で分離するのが難しいのではないかということが考えられる. 人手でも blog の個性から 20 代であるか 30 代であるかを判断するのは至難だろう.

提案手法は blog の個性を捉える手法であるので, 今回の例のように個性が似通っていると思われるクラス同士の分類では, ラベル無し事例の影響が大きくなると結果が悪化してしまうものと考えられる. これに対しては, ラベル無し事例の影響の強さをコントロールするパラメータを導入する事で回避できるのではないかと考えている.

7 結論と今後の課題

本研究では, blogを対象とする, ライティングスタイルやトピック傾向といったblogの個性に着目した半教師有り学習による分類方法を提案し, 評価実験によって提案手法の有効性を確認した. 提案手法は従来の半教師有り学習による文書分類と異なり, 複数のエントリから構成されるというblogの構造と, 同じblogに含まれるエントリは個性を共有するという性質を積極的に取り込んだ手法である.

今後の課題としては, まず副分類器の構築法, 選択法が挙げられる. 本稿ではランダムに2つのblogを選択するという単純な手法を用いたが, より効果的な選択法など, 研究の余地が残されている. また, 副分類器に用いる素性についてもさらに考慮する必要がある. 本稿で用いたbag-of-wordsではライティングスタイルを明確に捉えているとは言い難い. その他にも, 既存の半教師有り学習による文書分類手法との組み合わせや, スパムフィルタリングやblogのトピック分類といった他のタスクでの実験についても考えていきたい.

参考文献

[1] 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. 2006.

[2] 大倉 務, 清水 伸幸, 中川 裕志. スケーラブルで汎用的なブログ著者属性推定手法. 情報処理学会, 第181回 自然言語処理研究会 (SIGNL-181), 2007.

[3] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online Passive-Aggressive Algorithms. In *Journal of Machine Learning Research*, Vol.7, Mar, pp.551–585, 2006.

[4] Norihito Yasuda, Tsutomu Hirao, Jun Suzuki, and Hideki Isozaki. Identifying Bloggers' Residential Areas. In *Proceedings of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, pp.231–236, 2006.

[5] Hong Qu, Andrea La Pietra, and Sarah Poon. Automated Blog Classification: Challenges and Pitfalls. In *Proceedings of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, pp.184–186, 2006.

[6] J. Schler, Moshe Koppel, S. Argamon and J. Pennebaker. Effects of Age and Gender on Blogging. In *Proceedings of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, pp.199–205, 2006.

[7] Pranam Kolari, Tim Finin, and Anupam Joshi. SVMs for the Blogosphere: Blog Identification and Splog Detection. In *Proceedings of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, pp.92–99, 2006.

[8] Rie Kubota Ando and Tong Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. In *Journal of Machine Learning Research*, Vol 6, pp.1817–1853, 2005.

[9] Rie Kubota Ando and Tong Zhang. A high-performance semi-supervised learning method for text chunking. In *Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics (ACL-2005)*, pp.1–9, 2005.

[10] Rie Kubota Ando. Applying alternating structure optimization to word sense disambiguation. In *Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL-2006)*, pp.77–84, 2006.

[11] Chang Liu, Hwee Tou Ng. Learning Predictive Structures for Semantic Role Labeling of NomBank. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-2007)*, pp.208–215, 2007.

[12] John Blitzer, Ryan McDonald, and Fernando Pereira. Domain Adaptation with Structural Correspondence Learning. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP-2006)*, pp.1166–1169, 2006.

[13] John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, Bollywood, Boom-boxes, and Blenders: Domain Adaptation for Sentiment Classification. In *Proceedings of Association of Computational Linguistics (ACL-2007)*, pp.440–447, 2007.

[14] Toshiaki Fujiki, Tomoyuki Nanno, Yasuhiro Suzuki, Manabu Okumura. Identification of Bursts in a Document Stream. In *Proceedings of First International Workshop on Knowledge Discovery in Data Streams*, pp.55-64, 2004.

[15] Yasuhiro Suzuki, Hiroya Takamura, Manabu Okumura. Application of Semi-supervised Learning to Evaluative Expression Classification. In *Proceedings of the 7th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2006)*, pp.502–513, 2006.

[16] Kamal Nigam, Andrew Mccallum, Sebastian Thrun, and Tom Mitchell. Text Classification from Labeled and Unlabeled Documents using EM. In *Journal of Machine Learning* Vol .39, No.2, pp. 103–134, 2000.

[17] Avrim Blum and Tom Mitchell. Combining Labeled and Unlabeled Data with Co-training. In *Proceedings of the Workshop on Computational Learning Theory (COLT-1998)*, pp. 92–100, 1998.