少数の正例とラベルなし事例を用いた半教師付き学習に関する一考察

1X13C117-6　水落 洋貴
指導教員　後藤 正幸

1　研究背景と目的

近年，電子文書データの大規模化に伴い，これらを自動分類する手法の重要性が高まっている．自動文書分類では，所属するカテゴリが既知である文書を用いて分類器を学習し，新たに入力される文書のカテゴリを推定する．この自動分類手法は，スパムメールフィルタリングや成人年齢認証などを代表とした情報抽出にも応用されており，これらはユーザが所望する文書 (以下，正例文書) を抽出する技術である [1]．例えば文書推薦では，ユーザが既に閲覧した比較的少数の文書のみを正例ラベルが付与されている文書とみなせる状況が考えられる．この状況において，正例か正例でない (以下，負例) かを判別する分類器を構築するためには，少数の正例文書と，ラベルが付与されていない多数の文書 (以下，ラベルなし文書) から二値分類器を学習する必要がある．そこで，半教師付き学習の枠組みを援用し，ラベルあり正例文書とラベルなし文書を用いて学習を行う手法として Positive Exampled Based Learning(以下，PEBL) が提案されている [2].

PEBL では，ラベルなし文書のうち，正例文書の特徴から距離が離れた文書 (以下，仮負例) を「負例」とみなす．そして,「仮負例の抽出」および「正例と抽出された仮負例から分類器の学習」を繰り返しながら学習を行う．この手法では，正例文書集合と負例文書集合の分布が重なっていない場合には当てはまりがよい．しかしながら，文書の推薦などを考えると，正例文書集合と負例文書集合の分布が部分的に重なり合っており，正例である文書が仮負例として抽出されるおそれがあるため，正例文書数に対して仮負例数が多い可能性が高い．この場合，SVM の識別境界はデータ数が少ない方に寄ることが知られており，PEBL では正例に対して近い分類境界が学習される．そのため，抽出された文書がユーザの目的に合致する割合 (適合率) は高いが，その抽出数は極めて少なくなってしまう．しかしながら，正例文書を抽出する際は適合率だけでなく，ユーザの目的に合致する文書を正しく抽出できた割合 (再現率) も求められている．

そこで，本研究では正例文書に対して適合率を重視した分類器と，再現率を重視した分類器の 2 つの分類器を学習し，これらを組み合わせることで対象文書から正例を分類する手法を提案する．さらに本研究では，実際の新聞記事データを用いた分類実験を行い，正例文書に対する分類精度の観点から従来手法と比較して本研究の提案手法の有効性を示す．

2　従来手法

2.1　Support Vector Machine(SVM)

SVM [3] は，分離超平面（分類境界）から最も近いデータまでの距離（マージン）を最大化するように識別関数を学習し，二値判別を行う手法である．マージン最大化によって汎化能力を高めるという特徴があり,「高次元特徴空間」,「文書ベクトルの点在性」といった文書分類問題の特性により引き起こされる過学習という問題に対し有効とされている．

いま，全文書中に出現する V 種類の異なり語を単語集合 $\mathcal{W} = \{w_1, w_2, \ldots, w_V\}$ と表現し，またそれぞれの文書における各単語の出現頻度ベクトルを $\boldsymbol{x} = \{x_1, x_2, \ldots, x_V\}^T$ とする．このとき，各ラベル集合の学習データを分離する識別関数を係数ベクトル $\boldsymbol{a} \in \mathcal{R}^V$，バイアス項 b を用いて，式 (1) で表す．

$$f(\boldsymbol{x}) = \boldsymbol{a}^T\boldsymbol{x} + b \tag{1}$$

式 (1) に対してマージンを最大にする $\boldsymbol{a}$ を求める．識別関数 $f(\boldsymbol{x})$ の出力は，入力文書 $\boldsymbol{x}$ から識別境界への距離であり $f(\boldsymbol{x}) > 0$ ならば正例，$f(\boldsymbol{x}) < 0$ ならば負例と分類する．

2.2　PEBL

いま，少数の正例文書集合を $\mathcal{P}$，多数のラベルなし文書集合を $\mathcal{U}$，分類対象文書を $\mathcal{T}$ とすると，$\mathcal{P}$ と $\mathcal{U}$ を学習用文書として用いて分類器を構築し，$\mathcal{P}$ と同じラベルと推定される文書を $\mathcal{T}$ から抽出する問題を考える．この分類問題を扱う手法として PEBL [2] が提案されている．PEBL とは,「ラベルなし文書集合からの仮負例の抽出」と「少数の正例と抽出された仮負例を加えた仮負例集合を用いた分類器の構築」を繰り返しながら，分類器を学習する手法である．最終的に得られた分類器により，分類対象文書 $\mathcal{T}$ から正例を抽出する．具体的には最初の仮負例を抽出するため，単語 w_v ごとに正例文書 $\mathcal{P}$ に対する χ^2 統計量 $\chi^2(w_v, \mathcal{P})$ を式 (2) によって算出する．

$$\chi^2(w_v, \mathcal{P}) = \frac{E(AD - CB)^2}{(A+C)(B+D)(A+B)(C+D)} \tag{2}$$

$where$

$A = f(w_v, \mathcal{P}), B = f(w_v, \mathcal{U}), C = f(\overline{w_v}, \mathcal{P}), D = f(\overline{w_v}, \mathcal{U}).$

ただし，E は全学習用文書数，$\overline{w_v}$ は w_v 以外の全単語，$f(w_v, \mathcal{P})$ は正例文書集合 $\mathcal{P}$ 内の単語 w_v の出現頻度とする．PEBL の学習アルゴリズムは以下のとおりである．

STEP1) $i = 1$ とする．全単語 $\mathcal{W}$ のうち χ^2 統計量の上位 X 語を含まないラベルなし文書を i 回目の仮負例集合 $\mathcal{N}_i$ とする．

STEP2) $\mathcal{P}$ と $\mathcal{N}_i$ に対して SVM を用いて二値分類器を学習する．

STEP3) STEP2 で得られた分類器により，$\mathcal{N}_i$ を除くラベルなし文書集合 $\{\mathcal{U} \backslash \mathcal{N}_i\}$ を分類する．

STEP4) 負例と分類された文書があるならばそれらを新たな仮負例として $\mathcal{N}_i$ に追加し，$\mathcal{N}_{i+1}$，$i = i + 1$ として STEP2 へ．そうでなければ STEP5 へ．

STEP5) $\mathcal{P}$ と $\mathcal{N}_i$ により二値分類器を学習し，$\mathcal{T}$ から，正例を抽出.

3　提案手法

3.1　概要

従来手法 [2] では，新たな仮負例が抽出されなくなるまで仮負例の抽出・追加を繰り返し，最終的な分類器を学習する．この手法は，正例文書集合と負例文書集合が離れている問題ではうまく機能すると考えられる．しかしながら，実際の問題では正例文書集合と負例文書集合の分布に重なりがある場合が多い．その上で，正例文書数に対して仮負例数が多い場合は，従来手法は正例文書集合と負例文書集合の重なった部分で分類器を作成するため，データ数が少ない正例文書集合

に近い識別境界が学習される．そのため，この分類器では正例の適合率が高い一方で再現率は低くなってしまう．しかし，正例文書の分類では，抽出した文書が正例である割合も重要であるが，ラベルなし文書集合の中からより多くの正例文書を抽出することも求められる場合も考えられる．そのため，適合率と再現率のバランスを考慮するため，適合率および再現率それぞれを重視した 2 つの分類器を学習し，組み合わせることで，適合率と再現率を同時に考慮した分類を行う手法を提案する．

3.2 分類器学習方法

提案手法では，適合率を重視した分類器を Hard 分類器，再現率を重視した分類器を Soft 分類器とし，2 つの分類器 Hard，Soft を構築する．

Hard 分類器 (適合率重視)

適合率が高い従来手法 [2] においても，正例文書集合と負例文書集合の分布に重なりがある場合には，正例文書側に分布している負例文書が仮負例として抽出されず，適合率が悪くなってしまうことが考えられる．そこで，学習文書中のラベルなし文書をすべて仮負例とみなし，SVM により識別境界の学習を行う．これにより，正例に対し近い識別境界を持つ Hard 分類器を構築する．

Soft 分類器 (再現率重視)

再現率を重視する分類器を構築するには，正例文書集合全体を包括するような識別境界を学習すればよい．そのためには，なるべく正例文書から距離が遠い文書を抽出するべきであると考えられる．一方で正例文書に対して近い識別境界を持つ Hard 分類器は，負例文書側に存在する正例文書を抽出することが出来ない．そのため，この識別境界から距離の遠い文書を仮負例として抽出し，新たに正例文書との識別境界を学習することで，再現率が高い分類器を構築できると考えられる．そこで，Hard 分類器を用いて分類境界から負例側への距離が最も遠い N 件の文書を仮負例とみなして学習し直すことで正例に対し遠い識別境界を持つ Soft 分類器を学習する．

以上より構築された，Hard 分類器および Soft 分類器による分類対象文書に対する SVM 出力値をそれぞれ d_{Hard}，d_{Soft} とし，対象文書に対してそれらの和を算出し，予め設定した閾値 α により正例文書を抽出する．すなわち，分類対象文書のうち，

$$d_{\mathrm{Hard}} + d_{\mathrm{Soft}} > \alpha \tag{3}$$

となる文書を正例文書として抽出する．α の値は適合率と再現率を調整するパラメータであり，$\alpha \gg 0$ のときは適合率を，$\alpha \ll 0$ は再現率をより重視する分類規則である．

4 実験

4.1 実験条件

2000 年度の読売新聞記事データを用い実験を行う．本研究では，このデータ中の 28 個のカテゴリを対象とし，文書にはこのうちいずれか 1 つのカテゴリが割り振られている．学習文書として 1 カテゴリあたり 100 件の計 2800 件，分類対象文書として 1 カテゴリあたり 50 件の計 1400 件の文書を用いる．また，提案手法で用いるパラメータは $\alpha = 0$，$N = 50$ とした．比較手法として，PEBL，Hard 分類器，Soft 分類器を用いる．正例文書は，28 カテゴリから 1 つカテゴリを選択し，学習文書集合内の 100 件中 70 件はラベルあり正例文書とする．各カテゴリを正例としたデータセットを全 28 セットを用意し，式 (4)〜(6) で正例文書に対する適合率，再現率，F 値を算出し，それぞれの平均値をモデルの評価指標とする．

$$\text{適合率} = \frac{\text{正例と正しく抽出された文書数}}{\text{正例として抽出された文書数}} \tag{4}$$

$$\text{再現率} = \frac{\text{正例と正しく抽出された文書数}}{\text{正例文書数}} \tag{5}$$

$$F \text{ 値} = \frac{2 \times \text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}} \tag{6}$$

4.2 実験結果と考察

実験結果を表 1，2 に示す．

表 1．適合率，再現率，F 値

	PEBL	Hard	Soft	提案手法
適合率	0.922	**0.928**	0.254	0.739
再現率	0.267	0.305	**0.809**	0.514
F 値	0.414	0.459	0.387	**0.606**

表 2．データ数

	正例	仮負例
PEBL	70	2708.70(平均値)
Hard	70	2730
Soft	70	50

表 1 より提案手法は，PEBL に比べ，F 値において優れていることが分かる．このことから提案手法は，適合率を重視した分類器と再現率を重視した分類器を組み合わせることで，比較的高い精度を保った上でより多くの正例文書を抽出するバランスのとれた手法であるといえる．

表 2 より，PEBL ではラベルなし文書の大半を仮負例として抽出しており，再現率の大幅な低下を招いている．また，Hard 分類器よりも再現率が低下していることから，仮負例抽出後のラベルなし文書に負例が残っていると考えられる．

適合率を高めた Hard 分類器，再現率を高めた Soft 分類器において，それぞれ手法の目的である指標で優れていることが確認できる．このことより，本手法は，ユーザの目的に応じて α を変化させることで，適合率と再現率のトレードオフをコントロール可能と考えられる．

5 まとめと今後の課題

適合率を重視した分類器と再現率を重視した分類器を組み合わせることで，適合率と再現率を共に考慮した手法を提案し，新聞データを用いた実験により，その有効性を示した．

今後の課題として，N を変化させて学習した更なる複数の識別境界を混合する手法や学習段階でラベルなし文書から負例のみならず，正例も加える手法の提案などが考えられる．

参考文献

[1] 麻生英樹, 津田宏治, 村田昇, “パターン認識と学習の統計学,” 岩波書店, 2003.

[2] H. YU, H. Han, and K.C.-C. Chang, “PEBL: Positive Example Based Learning for web page classification using SVM,” Proc. ACM Special Interest Group on Knowledge discovery and Data Mining, pp. 239–248, 2002.

[3] C.Cortes and V. Vapnik, “Support-Vector Networks,” Machine Learning, Vol. 20, No.3 pp. 273–297, 1995.