評価文に対する二極指標の自動付与

岡野原 大輔† 辻井 潤一§‡
†東京大学理学部情報科学科 ‡CREST, 科学技術振興事業団
§東京大学大学院情報理工学系研究科コンピュータ科学専攻
{hillbig, tsujii}@is.s.u-tokyo.ac.jp

1 はじめに

近年，感想や意見など主観的な内容が入った文章（以下レビュー）を，それが評価対象を良くとらえているか（positive），悪くとらえているか（negative）で分類する評判分類の研究が盛んである [1, 2, 3, 4].

本稿ではこのタスクを拡張し，レビューを単に positive，negative で二値分類するのではなく，どの程度評価しているのかを指標として表す新しいタスクを提案する．我々はこの二極を持った指標を *sentiment polarity score* (sp-score) と呼ぶ．例えば，次の二つのレビュー「この本は最高です．かなりお薦めです」と「本の内容は悪くはないです」は従来の評判分類では共に positive に分類され，前者の方が強く薦めているという情報は失われてしまうが，我々のタスクでは，評価の程度を前者の sp-score は 4.6，後者の sp-score は 3.5 と表すことができ，比較することができる．

本稿では，Support Vector Regression (SVR)[5] を用いてこの指標を予測する手法を提案する．提案したタスクは**順序付多クラス分類**，つまり各 sp-score の値それぞれが別のクラスであり，各クラス間には順序関係が与えられている分類となっている．SVR はこの順序関係とレビューの関係を回帰を用いて学習する．また，SVR の回帰係数の成分値を調べることで表現や句単位での sp-score への影響力が求められることを示す．これに加え，評価分類では，個々の単語を素性に用いた，いわゆる bag-of-words ではとらえられない特徴も多いことから，本稿では N-gram や，評価対象の名詞周辺に現れる単語といった比較的単純だが bag-of-words より複雑な素性の効果についても述べる．

2 二極指標を用いた評価文分析

2.1 タスク設定

我々は評判分類を拡張し，どの程度評価しているのかを指標で表すタスクを提案する．この二極（positive, negative）を持った指標を *sentiment polarity score*（sp-score）と呼ぶ[1]. このタスクでは，sp-score が付けられたレビューを用いて学習し，新しく与えられたレビューの sp-score を正しく推測することが目標である．客観的にレビューの sp-score を測ることは難しいので，レビュアーが付けた sp-score を正解だとし，それを用いて分類器の性能評価を行う．以下，レビューから sp-score への写像である分類器を h，レビューと sp-score の組を (x, y) とする．テストデータ $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ が与えられた時，分類器 h の性能の評価尺度としては平均二乗誤差:$sq(h) = \frac{1}{n}\sum_{i=1}^{n}(h(x_i) - y_i)^2$ を用いる．これは大きな順序関係の間違いに対し，より大きなペナルティを与えるためである．

2.2 予備実験: 人による sp-score の予測

レビューの内容のみから，レビュアーがつけた sp-score を予測できるかは自明でない問題である．我々はこれを調べるため，被験者二人に対しそれぞれ，amazon.com[2]の本に関するレビューの sp-score を推測してもらった．各レビューにはレビュアーが本の評価に応じて 5 段階の評価を星の数として与えており，我々は，この星の数を sp-score の正解として用いた．

最初に被験者に対し，訓練データとして，sp-score が付けられた 20 レビューを見てもらい，レビューの内容と sp-score の関係を学習してもらった．次に，テストデータとして，sp-score が付けられていない 100 レビューを与え，それぞれ sp-score を付けてもらった．訓練，テストデータ共に各 sp-score が付けられたレビューの数は等しくしてある．

表 1 にその結果を示す．Random の行は sp-score をランダムに付けた場合，All3 の行は全て 3 をつけた場合の結果である．両被験者の結果は Random，All3 の場合に比べどちらとも有意差があり，レビューの内容のみから sp-score が予測できるといえる．

[1] 5 つ星による評価や，百点中の点数による評価が sp-score の一例として挙げられる．
[2] http://www.amazon.com

表 1: レビューに対する sp-score の推測結果．Random の行は sp-score をランダムに付けた場合，All3 の行には全て 3 をつけた場合の結果．

	二乗誤差
Random	3.20
All3	2.00
被験者 1	0.77
被験者 2	0.79
被験者の平均	0.78

3 機械学習を用いた二極指標の付与

本章では機械学習を用いた二極指標の付与について述べる．最初に 3.1 でレビューを特徴ベクトルで表現する方法を述べ，3.2，3.3，3.4 で特徴空間からの sp-score への写像を学習する方法を述べる．また素性の工夫について 3.5 で述べる．

3.1 特徴ベクトル

一般の文書分類や情報検索と同様に我々はレビューを実数値の多次元ベクトルである特徴ベクトル $x = (x_1, x_2, x_3, ..., x_l)$ で表現する [6]．ベクトルの各要素は，レビューから得られる素性である．我々は素性の重みとして tfidf を用いた．tfidf は次の通り定義される:$\mathrm{tfidf}(w,d) := \mathrm{tf}(w,d) * \log \frac{D}{\mathrm{df}(w)}$．ただし，$\mathrm{tf}(w,d)$ は文書 d 中の単語 w の出現回数，$\mathrm{df}(w)$ は単語 w を含む文書数，D は全文書数である．また，長いレビューと短いレビューの特徴ベクトルを同じスケールで扱えるように，得られた特徴ベクトルに対し L_2 ノルムが 1 となるように正規化を行った．

3.2 sp-score への写像の学習

レビューの sp-score を予測する問題は順序付多クラス分類問題の一種だと考えられる．つまり，sp-score の値それぞれが別のクラスであり，各クラス間には順序関係が与えれている．予測を間違った際のペナルティは，多クラス分類のように全て等しいのではなく，二乗誤差として順序関係の大きな間違いに対し大きなペナルティが与えられる．

この問題は順序関係を無視することで，一般的な多クラス問題の枠組みで解くこともできる．我々は pairwise Support Vector Machines (pSVM)[7] を多クラス分類器として利用した．

それに対し順序関係を利用して特徴ベクトルから sp-score への写像を学習するのに，我々は分類ではなく回帰を用いる．最小二乗法により回帰を求める手法は，文書が通常数千から数万次元の特徴ベクトルを持つことから過学習を起こしやすい．そのため我々は過学習をコントロールできる Support Vector Regression (SVR)[5] を用いる．

それぞれの説明の前に pSVM，SVR 両方の基となる Support Vector Machine(SVM) について簡単な説明を行う．SVM は，二値分類を解く学習器であり，線形関数 $f(x) = \langle w \cdot x \rangle + b$ を識別関数として用いる．SVM はこの訓練サンプルを最も大きなマージンで分類する平面を求める．また，SVM は写像された特徴空間をカーネル関数を用いて陰に扱うことができるため，線形分離できない識別問題にも適用できることが知られている [8].

3.3 多クラス分類

pSVM[7] は，多クラス分類問題を解くための分類器である．m 個のクラスを分類する時は $m \cdot (m-1)/2$ 個の SVM による分類器を用意し，それぞれが全てのクラスの対について独立に二値分類の学習を行う．そして，全体の分類器の出力は，各分類器の結果の多数決により最も多く選ばれたクラスとする．

pSVM を sp-score の予測に用いる際，sp-score はそれぞれ独立の異なるクラスだとされ学習される．例えば，sp-score が $\{1,2,3,4\}$ の場合は 6 個の分類器を用意し，それぞれが 1vs2, 1vs3, 1vs4, 2vs3, 2vs4, ,3vs4 の二値分類の学習を行う．もし与えられたレビューに対する各分類器の結果が $1,1,1,2,4,3$ ならばレビューの sp-score は 1 と推定される．

pSVM による予測は，各クラスに順序関係があることを無視しているため，大きな間違いを起こしやすいと予想される．また，sp-score の粒度が細かくなるにつれて，各二値分類器が利用する訓練データがスパースになる問題を抱えている．

3.4 SVR による学習

SVR[5] は訓練データから回帰を求めるための学習器であり，SVM と同様に最大マージン原理，カーネル関数を利用することで高い汎化能力を備えつつ，入力を高次元に写像した上で回帰を求めることができる．

SVR を用いて sp-score を予測する場合，入力を特徴ベクトル，出力を sp-score とした回帰問題としてとらえ，訓練データから回帰平面を求める．

SVR を用いる利点として，順序関係を自然に学習に用いることができることや，sp-score の粒度が細かくなった場合にも訓練データがスパースにならない点があげられる．しかし，SVR はカーネル関数を用いることがで

きるものの，回帰が求まらない場合がある．例えば，単純な例として，特徴空間が 1 次元であり，訓練データが $(x = 1, y = 1)$ $(x = 2, y = 2)$, $(x = 3, y = 8)$ の場合，特徴空間を非線形にスケーリングしなければこの訓練データに対する回帰は学習できない．そのため SVR による推定では中央の値は正しく推定できるものの，端の値では推定は難しくなると考えられる．

SVR は pSVM と違い，それぞれの素性がどのように sp-score の決定に寄与しているかを回帰係数 w の成分値として直接得ることができる点も大きな特徴である．（ただし，カーネル関数が線形カーネルの時）．これより，どのような句や表現が sp-score に影響したかを数値として得ることができる．

3.5 素性の工夫

本タスクでは二値分類ではなく sp-score を測るため，程度を表す副詞と感情や意見を表す形容詞との組み合わせが重要となる．そのため従来の単語単位の素性，unigram に加え，二単語の連なりからなる bigram と三単語の連なりからなる trigram を導入した．例えば “very good” という表現は bigram の素性により初めてとらえられる．

また，評価対象の名詞（句）周辺に出現した素性は評価決定に重要だと考えられることから，我々は次の二種類の素性を用いた．“inbook” は，book と同じ文中に出現した単語により発火する素性であり，“abook”(around book) は book の両隣二単語以内に出現した単語により発火する素性である．本についてのレビューであれば，“book” が出現した文，もしくは周辺単語に評価決定に重要な情報（例　*This book is good*) が高く出現する可能性がありこれらの素性が精度を向上させる可能性がある．Mullen[9] らが重要単語 (e.g. book) を参照している単語を人手で決定し素性に加えたのに対し，我々の手法は，周辺，文単位で重要単語を参照していると思われる単語を自動で素性として加えた形となっている．

4 実験

4.1 実験設定

実験は 2.2 で用いたデータと同じデータを利用した．前もって，レビュー中の全ての HTML タグ及び，句読点は取り除き，ステミング処理を行った．このデータから二種類のテストセットを作成した．一つ目は Harry Potter の本に関する 1650[3] レビューからなる corpus A であり，二つ目は様々な本に関する 1250 レビューからなる corpus B である．2.2 で用いたデータは，corpus A から選ばれた 100 レビューであり，スケールは違うものの corpus A の結果と 2.2 の人による結果は比較可能である．corpusA，B の 1 レビュー当たりの平均単語数はそれぞれ 165.0，112.8，平均文数はそれぞれ 9.3，5.7 であった．sp-score が違うレビューの間に，文数，単語数の大きな差は見られなかった．訓練，テストデータ共に各 sp-score が付けられたレビューの数は等しくしてある．SVM 及び SVR の学習に *SVMLight*[4] を用いた．多項式 2 次カーネルを使用し，コストパラメータ $C = 100$ を用いた．以下の実験は全て 10-fold の交差検定の結果である．

表 2: pSVM と SVR の比較．どちらも素性に unigram/tfidf を用いている．値は平均二乗誤差を表す．

手法	Corpus A	Corpus B
pSVM	1.32	2.13
SVR	**0.94**	**1.38**

表 3: N-gram や重要単語の周辺に出現した素性を用いた結果．値は平均二乗誤差を表す．

素性	Corpus A	Corpus B
unigram	0.94	1.38
+ bigram	**0.89**	1.41
+ trigram	0.90	1.42
+ inbook	0.97	**1.36**
+ abook	0.93	1.37

4.2 pSVM と SVR との比較

表 2 に，pSVM と SVR を用いて sp-score を推定した時の結果を示した．どちらも素性は unigram，重み付けは tfidf を用いた．この結果から，どちらの Corpus においても，SVR の方が二乗誤差で優れていることがわかる．これは，3.2 で述べたように SVR は順序関係を利用して学習できているのに対し，pSVM は順序を無視して学習しているためである．また，この結果からは sp-score は，複雑な意味の重ね合わせではなく，単純な特徴ベクトルからの回帰によって，求められる可能性を示唆している．

SVR は 3.4 で指摘したように，非線形の回帰が行えない問題点がある．また訓練データが離散的であることも用いていない．更なる性能向上のためには，出力に構造がある場合を取り扱える学習器 [10] などさらにこのタス

[3] 訓練データ数を変えて得られた学習曲線によると，訓練データ数を増やすことで精度はまだ向上することが言える．

[4] http://svmlight.joachims.org/

表 4: SVR の w の成分値の絶対値が大きかった部分に対応する素性．value の行は回帰係数における成分値を表す

positive		negative	
value	bigram	value	bigram
1.73	best book	-1.61	at all
1.69	is a	-1.50	waste of
1.49	reat it	-1.38	potter book
1.44	all age	-1.36	out of
1.30	can't wait	-1.28	not interest

クに特化した学習器の導入が有効であると考えられる．

4.3 素性の比較

分類に用いる素性が unigram のみの場合（ベースライン）に対し，3.5 で述べた素性を加えた時の精度について調べた．いずれも素性の重み付けに tfidf，分類器に SVR を用いている．表 3 に各素性を unigram に加えた時の平均二乗誤差を示す．この結果からは，Corpus A では，単純な bigram や trigram が精度の向上に貢献している一方で評価対象の周辺単語を使った素性（inbook, abook）の有効性はわずかであるという結果が得られた．この原因として，評価対象の単語は頻出し，新たに加えた素性集合が従来の unigram の素性集合とほとんど変わらなかった点が挙げられる．その一方で Corpus B では，bigram，trigram を用いたことにより精度は低下した．Corpus B では，1 レビュー当たりの単語数が少なく，またジャンルが多岐に渡るため，頻出する bigram が Corpus A と比較し少なく有効に働かなかったためだと思われる．

4.4 有効な素性の抽出

線形カーネルの SVR を用いて corpus A の訓練データを学習した際の回帰係数 w の成分値が大きかった素性のうち，bigram であるものを表 4 に挙げる．Corpus B では "highly recommend"(value = 1.19)，"great !" (value = 1.14)，"very disappoint" (value = -1.19) 等が高い成分値を持っていた．結果からは，評価決定に重要であると思われる句が高い順位にあり，データマイニングなどに使える可能性を示している．

しかし，この成分値をそのまま素性単位の評価指標として利用するには，まだ多くの課題が残されている．一つ目に回帰係数の成分値は，レビュー中に共起する他の素性の影響も受けているため独立に決定されていないこと，二つ目に表 4 中の " is a" のように評価決定に寄与していないと思われる素性が訓練データの偏りにより含まれている可能性があることが挙げられる．今後，素性選択や，客観的な情報の排除などの処理が必要であると考えられる．

5 結論

本稿では，レビューが評価対象をどの程度評価しているのかを二極指標 (sp-score) を用いて予測する新しいタスクを提案し，それに対する解法として Support Vector Regression (SVR) を用いる手法を提案した．また，N-gram や評価対象名詞句の周辺の単語を素性として用いる手法による性能向上を確認した．我々の提案した手法による最良の結果は最小二乗誤差で 0.89 であり，人の結果 0.78 と比べても十分正確に予測できているといえる．また，SVR の回帰係数の成分値を調べることにより，どの素性が sp-score の決定にどれだけ寄与しているかが数値として得られることを示した．今後，非線形の回帰や出力が構造を持った場合の学習 [10] を行うことで性能向上を目指す他，データマイニングとしての用途の可能性を調べる予定である．

参考文献

[1] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proc. of EMNLP*, pages 79–86, 2002.

[2] B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proc. of ACL*, 2004.

[3] T. Kudo and Y. Matsumoto. A boosting algorithm for classification of semi-structured text. In *Proc. of EMNLP*, pages 301–308, 2004.

[4] P. D. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proc. of ACL*, pages 417–424, 2002.

[5] A. Smola and B. Sch. A tutorial on Support Vector Regression. Technical report, NeuroCOLT2 Technical Report NC2-TR-1998-030, 1998.

[6] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[7] U. Kresel. *Pairwise Classification and Support Vector Machines Methods*. MIT Press, 1999.

[8] J. S. Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambiridge Univsity Press, 2004.

[9] A. Mullen and N. Collier. Sentiment analysis using Support Vector Machines with diverse information sources. In *Proc. of ACL*, 2004.

[10] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *Proc. of ICML*, 2004.