複数の分類結果の信頼度を利用したレビュー記事の自動分類
Sentiment classification using reliability of multiple classification results

堤 公孝 † **嶋田 和孝** ‡ **遠藤 勉** ‡

Kimitaka Tsutsumi Kazutaka Shimada Tsutomu Endo

† 九州工業大学大学院 情報工学研究科 情報科学専攻

Graduate School of Computer Science and Systems Engineering,Infomation Science, Kyushu Institute of Technology

‡ 九州工業大学大学 情報工学部 知能情報工学科

Department of Artificial Intelligence, Kyushu Institute of Technology

Abstract: Customers often consult the WWW for the product information before buying a product. Bulletin boards on online-shopping sites usually include opinions for products. More precise and effective methods for evaluating the products are useful for customers.

In this paper, we propose a method to classify documents into positive or negative opinion. The method consists of two classifiers based on SVMs and score calculation. First we improve the accuracy of a scoring method by using a feature selection process. Then we explain classification rules based on the two classifiers. The proposed method classifies documents on the basis of confidence of the results of each classifier.

We compare the proposed method with related work. Experimental results show the effectiveness of our method.

1 はじめに

近年，我々はインターネットなどを通じて，レビュー記事を大量に入手する事ができる．そして，その記事を製品購入などの意思決定の際に参考にしている．しかし，大量のレビュー記事の内容を把握するのは容易ではなく，内容を把握するために，記事の効率的な活用法が必要となる．我々がレビューを参考にする場合，まず，その情報が肯定的意見なのか，否定的意見なのかを判断して参考にしている．レビュー記事を肯定/否定的意見に自動分類することができれば，内容把握が容易になり，情報をより有効に活用できる．例えば，悪い評判が少なく良い評判が多い製品は，良い製品なのではないかといった分析ができる．本研究では，レビュー記事を肯定/否定的意見に自動分類することを目的としている．

現在，レビュー記事を肯定的/否定的意見に自動分類する研究は盛んに行われている [1]．Pang ら [2] は，bag-of-words のみを素性として用いて，ナイーブベイズ法 (Naive Bayes)，最大エントロピー法 (Maximum Entropy)，サポートベクタマシン (SVM) などの機械学習器によって分類する手法を提案した．英文で書かれた映画レビュー記事を対象として実験を行い，機械学習器の中でも，SVM において最も高精度で分類を行えることを示している．筏島ら [3] は，肯定/否定的評判で素性が出現する割合を用いたスコアリングにより文単位での分類を行っている．藤村ら [4][5] は，肯定/否定的評判での素性の出現確率を用いたスコアリングにより分類することを提案している．筏島ら，藤村らはこの手法を用いることで，SVM と同等，または SVM より高い精度が得られたとしている．この他にも，共起情報を用いた分類 (Turney ら [6])，語の系列パターンや部分依存木を素性として用いた分類 (Matsumoto

九州工業大学 情報工学部 知能情報工学科
〒 813-0036
飯塚市川津 680-4
E-mail:{k_tsutsumi, shimada, endo}@pluto.ai.kyutech.ac.jp

ら [7])，意見文のみを抽出し分類に用いる分類法 (Pang ら [8]) など，様々な分類法が提案されている．

現在提案されている手法は，1 つの分類器を用いて分類を行っている．複数の分類器の分類結果を利用し，誤分類している部分を補い合う事で，より良い精度が得られる可能性がある．本研究では，Pang らが提案した SVM を用いた分類と，筏島らが提案したスコアリング手法による分類の，2 つの分類結果の信頼度を利用して分類を行う手法を提案し，その分類精度を調査する．

以下，2 章では用いる分類器について述べ，3 章では提案手法について説明する．続いて，4 章では今回の手法の実験について述べ，考察する．最後に，5 章でまとめる．

2 分類器

本研究では，従来の手法である，SVM を用いた分類手法，スコアリングを用いた分類手法を組み合わせて用いる事で分類を行う．ここでは本研究で用いた分類器として，Pang らが用いた SVM による分類手法，筏島らが用いたスコアリングによる分類手法について説明し，それぞれの分類器で分類実験を行い，精度を確認する．分類実験には，Pang らが用いたものと同じ英文の映画レビューデータセットを用いる．最後に，各分類器でのエラー分析を行い，本研究での提案手法を考案した動機について述べる．

2.1 SVM を用いた分類

SVM は Vapnik らが考案した機械学習器であり，最も良い分類精度を得ている学習器の 1 つとして知られている [9]．図 1 に SVM の学習モデルを示す．訓練データを素性のベクトルの集合で表し，訓練データ集合の中で，境界付近に存在する訓練データであるサポートベクターと境界線との距離であるマージンを最大化するように分離超平面 (hyperplane) を決める．

Pang らは，単語を素性として用いており，その単語がそのドキュメントに存在すれば 1，存在しなければ 0 として学習を行うことで最も高い精度を得ている．

本研究でも，Pang らと同じ手法を用いて SVM による分類実験を行い，精度を調査した．SVM は，ツールとして SVM ^{light} *を使用し，線形カーネルを用いて実験を行った．実験の結果，82.1%の精度を得た†．

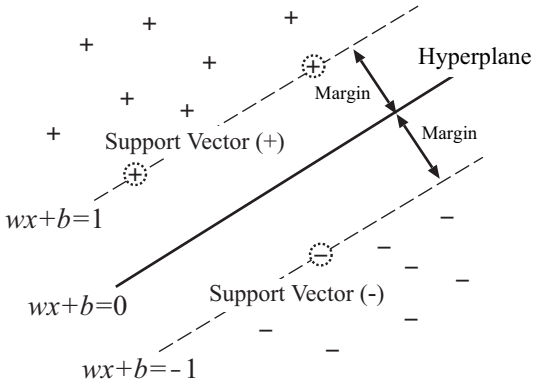

図 1: SVM の学習モデル

2.2 スコアリング手法を用いた分類

筏島ら [3] は，肯定的な評判には肯定的な概念を持った素性が，否定的な評判には否定的な概念を持った素性が偏って存在するはずであるという仮定の基，スコアリングをして分類を行っている．$pos(w_i)$ と $neg(w_i)$ は，それぞれ，素性 w_i が肯定もしくは，否定的評判に含まれる数を表している．また，$\sum pos$ と $\sum neg$ は，肯定的評判に含まれる素性の総数と否定的評判に含まれる素性の総数である．式 1 は，素性 w_i が肯定もしくは，否定的評判にどれくらい偏って出現するかを表しており，素性 w_i が肯定的評判に偏って出現すれば正の値，否定的評判に偏って出現すれば負の値になる．この偏りを利用した素性 w_i のスコアを $Score_{IDF}(w_i)$ としている．

$$Score_{IDF}(w_i) = log\left(\frac{pos(w_i)+1}{\sum pos} \times \frac{\sum neg}{neg(w_i)+1}\right) \quad (1)$$

分類の際には，文書中の語のスコアの総和 (式 2) を求め，その値が正の値なら肯定的評価の文書，負の値なら否定的評価の文書としている (式 3)．

*http://svmlight.joachims.org/

†Pang らの先行研究での正解率 82.9%と本研究での結果が異なるのは，前処理の素性選択を先行研究と完全に一致させる事ができなかったためであると考えられる．前処理については，4.1 を参照のこと．

$$Score(d) = \sum_{ALL w_i \in d} Score_{IDF}(w_i) \tag{2}$$

$$d = \begin{cases} Positive & (Score(d) > 0) \\ Negative & (Score(d) \leq 0) \end{cases} \tag{3}$$

本研究では，スコアリング手法を用いた分類に対して，品詞情報の利用と，χ^2 検定による素性の選別という 2 つの拡張を行っている．

- 品詞情報の利用

  評価文を扱った関連研究の多くでは，品詞情報を用いている．そして，形容詞は，最も直接的に評価を表す品詞である．本研究では，形容詞のスコアを大きくする事で精度の向上を図っている．具体的な処理を以下に記す．

  まず，元のデータセットのそれぞれの単語に対して品詞タグを付与する．品詞タグの付与には，brill's Tagger [‡]を用いている．そして，分類の際には，訓練データの中で 3 回以上形容詞タグを付与されている素性の重みを定数倍している．ここで 3 回以上としたのは，タグ付与誤りの影響を小さくするためである．形容詞の重みは経験的に求め，一番精度が良かった 7 倍を採用している．

- χ^2 検定による素性の選別

  藤村ら [5] は，χ^2 検定を用いて，その素性がどれだけ偏って現れるかを数値化し，信頼性の高い素性を選別する手法を提案している．χ^2 値の求め方は式 4 に示す．χ^2 値が大きい素性ほど，統計的に信頼性が高く，肯定/否定の文書に偏って存在する．そこで，信頼性が低い素性，つまり χ^2 値が小さい素性を棄却し，信頼性が高い素性のみを分類に用いる．

$$\chi^2(w_i) = y_{w_i} \cdot \sum_{w,class} \frac{(df(w,class) - df_{total} \cdot p_w \cdot p_{class})^2}{df_{total} \cdot p_w \cdot p_{class}}$$

$$w \in \{w_i, \bar{w_i}\}$$

$$class \in \{positive, negative\} \tag{4}$$

[‡]http://www.cs.jhu.edu/~brill/home.html

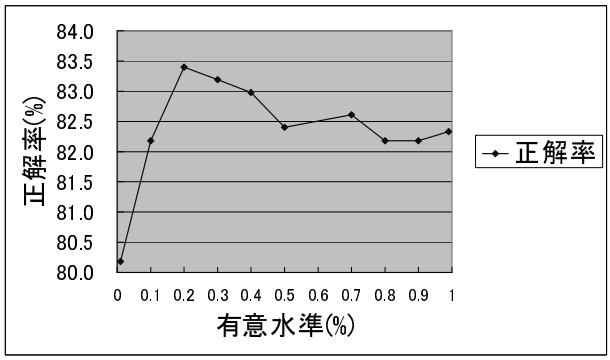

図 2: 有意水準と正解率

表 1: 拡張前後での分類精度

分類方法	正解率
拡張前	79.8%
拡張後	83.4%

正解率が最高となる有意水準を求めるための実験を行った．実験結果を図 2 に示す．結果より，有意水準が 20%のときに精度が最高になった．したがって，今回は有意水準 20%を採用した．

拡張を行う前と後での分類精度を調査するために実験を行った．結果を表 1 に示す．結果より，品詞情報の利用，χ^2 検定による素性の選別の 2 つの拡張を行う事で，拡張前よりも高い精度を得る事ができた．

2.3 エラー分析

誤分類文書がどのようなスコア，分離平面からの距離で分布しているかを調査した．スコアリング手法での分布を図 3 に，SVM での分布を図 4 に示す．結果より，どちらの分類器を用いた場合でも境界線に近いほど誤分類している文書数が多く，境界線付近での分類結果の信頼性が低い事を表している．

誤分類文書総数はそれぞれ，スコアリング手法による分類では 232 個，SVM による分類では 251 個であった．どちらの分類手法でも同じ文書を間違えていたのは 119 個であり，誤分類したもののうち，同じ文書を間違えて分類しているのは約半分の文書であることが

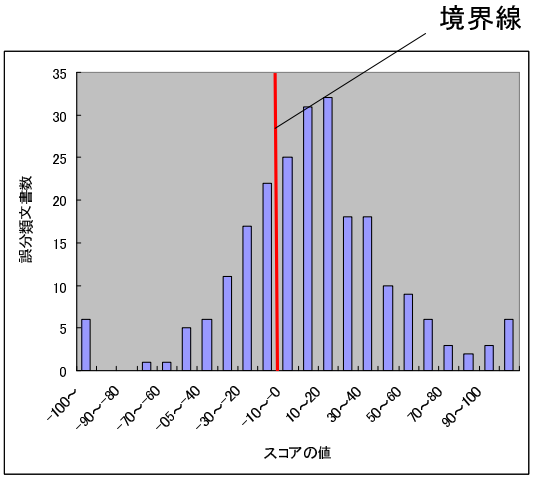

図 3: スコアリング手法での誤分類文書の分布

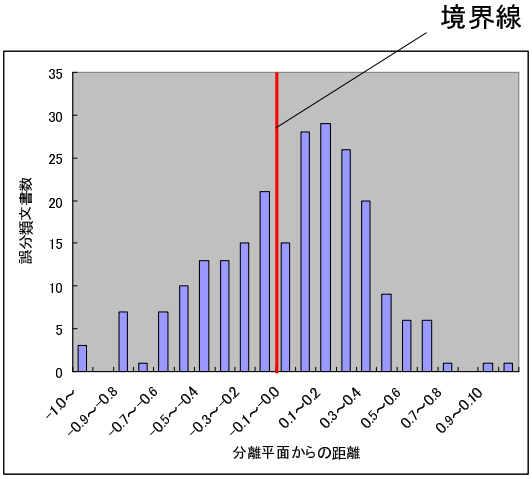

図 4: SVM での誤分類文書の分布

分かる．したがって，一方の分類器では誤って分類していても，もう一方の分類器では正しく分類できている文書が多数存在しており，互いの誤分類を補い合う事で精度が向上する可能性がある事を示唆している．

3 提案手法

前節のエラー分析の結果より，境界線に近いほど分類の信頼性が低く，2 つの分類器で誤分類を補い合う事で精度が向上する可能性がある事が分かった．境界線からの距離は，SVM の場合は分離平面からの距離，スコアリング手法の場合はスコアの絶対値に相当する．文献 [4] でも，スコアの絶対値が大きい文書ほど分類精度が高いと報告している．本研究では複数の分類器を用いて分類を行い，より境界線からの距離が大きい，信頼性が高い分類結果を採用する手法を提案する．

3.1 複数の分類器を用いた分類手法

本研究では，前節で説明した，SVM とスコアリングに基づく分類器を組み合わせて分類を行う．ここでは，2 つの分類器を用いて分類を行う手法として，3 つのタイプの手法を提案する．具体的な手順を以下に記す．k, l, m はある定数とする．

タイプ (1) スコアリング手法での分類結果の信頼度に着目する場合

条件： スコアの絶対値が k 以下 かつ 分離平面からの距離が l 以上

条件を満たす
SVM での分類結果を最終的な分類結果にする

条件を満たさない
スコアリング手法での分類結果を最終的な分類結果にする

タイプ (2) SVM での分類結果の信頼度に着目する場合

条件： 分離平面からの距離が l 以下 かつ スコアの絶対値が k 以上

条件を満たす
スコアリング手法での分類結果を最終的な分類結果にする

条件を満たさない
SVM での分類結果を最終的な分類結果にする

タイプ (3) 片方の分類器に依存せず，両方の分類結果に着目する場合

条件： スコアの絶対値 $<$ 分離平面からの距離 $\times m$

条件を満たす
SVM での分類結果を最終的な分類結果にする

条件を満たさない
スコアリング手法での分類結果を最終的な分類結果にする

タイプ (1) は，スコアリング手法による分類結果の信頼度に着目し，信頼度が低ければ SVM による分類結果を用いて分類器を組み合わせて分類を行う方法である．逆に，タイプ (2) は SVM による分類結果の信

頼度に着目し，信頼度が低ければスコアリング手法による分類結果を用いて分類器を組み合わせて分類を行う方法である．タイプ (3) は，どちらかの分類器の結果の信頼度に着目するわけではなく，それぞれの分類器での信頼度の高さを比べて，信頼度が高いほうを採用して分類を行う方法である．

4 実験

本研究では，Pang ら [2] が用いたものと同じ英文の映画レビューデータセット§を対象として，3 分割交差検定により実験を行っている．このデータセットは，肯定及び否定のレビューそれぞれ 700 本，計 1400 本で構成されている．素性は，Bag-of-words 素性として，全レビュー中で出現回数 4 以上の単語を用いた．ステミング処理や，ストップワードを用いた語の削除は行っていない．

4.1 前処理

“ not, don’t ”などの否定の意味を持つ語は，その語の後の動詞，形容詞などの意味を反転させる語である．例えば，“ It is not good . ”という文の場合，good は肯定の意味を表す語と考えられるが，not があるため意味が反転し，否定の意味を表す．こういった表現を無視すると，分類精度に悪影響を及ぼすと考えられる．Pang らは否定文に対する前処理を行っており，本研究でも同じ手法を用いた．具体的な処理としては，まず，否定の意味を持つ語を含む文を見つけ，否定の意味を持つ語から，区切り文字 (コンマ，ピリオドなど) までの語全てに対して,「NOT_」というタグをつける．処理の例を以下に示す．

否定文の前処理の例

文: It is not good .

処理後の文: It is not NOT_good .

このように，意味が反転したことが分かるようにタグを付与することで，本来の意味とは逆の意味を持ってしまう素性に対し，そのまま，その語の意味での重みをつけてしまうことを防いでいる．本研究では，否定の意味を持つ語として，“ **not, never, no, nobody, nothing, none** ” を用いた．

§http://www.cs.cornell.edu/people/pabo/movie-review-data/

表 2: 複数の分類器を用いた分類

分類方法	正解率
タイプ (1)	85.3%
タイプ (2)	83.3%
タイプ (3)	85.2%
スコアリング	83.4%
SVM	82.1%

4.2 評価実験

タイプ (1)，タイプ (2)，タイプ (3) のそれぞれの手法を用いて評価実験を行った．k, l, m の値は経験的に求め，それぞれ最適な値である，20，0.2，50 とした．実験結果を表 2 に示す．

まず，複数の分類器を組み合わせて分類を行った手法は全て，単分類器による分類である，スコアリング手法による分類，SVM による分類と同等かそれ以上の精度を得ている．タイプ (2) の精度は 83.3%で，スコアリング手法による分類の精度の 83.4%に比べて若干劣っているが，これは，タイプ (2) が SVM の分類結果の信頼度に着目して，SVM の分類精度を底上げしようとしている手法であるためであり，SVM の精度 82.1%に比べると 1%程度の精度向上が見られる．したがって，複数の分類器を利用した本手法を用いることによって，単分類器を用いた場合よりも精度が向上する事を確認できた．

次に，タイプ (1)，タイプ (2) では，分類精度が高い手法であるスコアリングでの分類結果に着目しているタイプ (1) の方が高い精度を得ている．これにより，2 つの分類器を組み合わせて用いる場合，分類精度が高い分類器の分類結果に着目して分類を行った方が，精度が高い事が分かる．これは，あらかじめ，組み合わせる分類手法のどちらの方が分類精度が高いか分かっていないと，高い精度を得る事ができないという問題点を含んでいる事を意味する．タイプ (3) は，タイプ (1) に比べて若干精度が劣るが，ほぼ同程度の精度を得ており，そのような問題点は無い．今回は，分離平面からの距離を単純に定数倍してスコアの絶対値と比べているが，2 つの値を正規化して比べる手法が考案できれば，さらに良い精度が得られる可能性がある．

また，今回は k, l, m の値は経験的に求めているが，

このままでは汎用性が低い．したがって，自動的に最適な値を推定する手法を検討する必要がある．

5 おわりに

本研究では，複数の分類結果の信頼度を用いたレビュー記事の自動分類手法を提案した．実験結果より，各分類手法を単体で用いるよりも，2 つの手法で分類を行い信頼性の高い手法の分類結果を採用する事で高い精度を得る事ができた．

今回は Bag-of-Words 素性のみを用いて分類を行っているが，系列パターン，部分依存木など，Bag-of-Words 以外の素性を Bag-of-Words と共に分類に用いる事を提案し，高い精度を得ている研究がある [3]，[7]．Matsumoto ら [7] は本研究と同じデータセットを用いて分類を行い，高い精度を得ている．また，Pang ら [8] は，意見文のみを抽出して分類に用いる手法を提案し，同じデータセットで実験を行い高い精度を得ている．このような，より高い分類精度の手法を組み合わせて本研究の手法を用いて分類を行う事で，さらなる精度向上の可能性がある．

Pang ら [10],Okanohara ら [11] のように，肯定/否定の 2 値分類ではなく，さらに細かい粒度で分類を行っている研究もある．より細かい粒度で分類を行う事で肯定/否定の度合いを知る事ができ，情報の有用性は増す．今回は，境界線からの距離を分類の信頼度として位置付けたが，境界線からの距離を肯定/否定の度合いとして位置付ける事で，このような多値分類にも対応させた手法に拡張できる可能性がある．

参考文献

[1] 乾 孝司，奥村 学：テキストを対象とした評価情報の分析に関する研究動向，自然言語処理 Vol.13, No.3, pp.201-242, 2006.

[2] Bo Pang and Lillian Lee and Shivakumar Vaithyanathan. Thumbs up? Sentiment Classification using Machine Learning Techniques. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp.79-86, 2002.

[3] 筱島郁子, 嶋田和孝, 遠藤勉：系列パターンを利用した評価表現の分類, 言語処理学会第 11 回年次大会, 2005.

[4] 藤村滋，豊田正史，喜連川優: 電子掲示板からの評価表現および評判情報の抽出，人工知能学会全国大会 (第 18 回),2004.

[5] 藤村滋，豊田正史，喜連川優：文の構造を考慮した評判抽出手法，電子情報通信学会データ工学ワークショップ (DEWS),2005.

[6] Peter D. Turney. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp.417-424, 2002.

[7] Shotaro Matsumoto and Hiroya Takamura and Manabu Okumura. Sentiment Classification using Word Sub-Sequences and Dependency Sub-Trees. Proceedings of the 9th Pacific-Asia International Conference on Knowledge Discovery and Data Mining (PAKDD-05), 2005.

[8] Bo Pang and Lillian Lee. A Sentiment Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL), pp.271-278, 2004.

[9] Vladimir Vapnik.(1999). Statistical Learning Theory. Wiley.

[10] Bo Pang and Lillian Lee. Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales. Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL), 2005.

[11] Daisuke Okanohara and Jun-ichi Tsujii. Assigning Polarity Scores to Reviews Using Machine Learning Techniques. Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP), pp.314-325, 2005.