

# 複数の分類結果に基づく映画レビューの極性判定

堤 公孝\*<sup>1</sup>      嶋田 和孝\*<sup>2</sup>      遠藤 勉\*<sup>2</sup>  
 Kimitaka Tsutsumi      Kazutaka Shimada      Tsutomu Endo

\*<sup>1</sup>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, Information Science, Kyushu Institute of Technology

\*<sup>2</sup>九州工業大学 情報工学部 知能情報工学科

Department of Artificial Intelligence, Kyushu Institute of Technology

In this paper, we propose a method to classify movie review documents into positive or negative opinions. The method consists of three classifiers based on SVMs, ME and score calculation. The proposed method classifies documents by using a voting process based on the results of each classifier. We compare two voting process for our method: naive voting and weighted voting. In addition, we compare the proposed method with related work. Our method consisting of 3 single classifiers improved the accuracy. The experiment result shows that our method contains a possibility that it boosts up the accuracy essentially.

## 1. はじめに

我々は、Web 上などのレビュー記事を、製品購入などの意思決定の際に参考にしていく。そして、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する。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。

現在、レビュー記事を肯定/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、文書単位の分類において、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ている。また、箆島ら [3] は、文書単位の分類において、単語のスコアリングに基づく分類手法を提案して実験を行い、SVM より高い精度を得ている。他にも、様々な分類手法や素性抽出法を利用することで精度が向上したという報告がなされている [4][5]。

しかしながら、従来の研究では、1つの分類器のみで分類を行っている。著者らはこれまでに、SVM と箆島らが提案したスコアリング手法を用いた分類結果を組み合わせた分類手法を提案し、個々の分類器を単体で用いた場合よりも高い精度が得られることを確認している [6]。本研究では、SVM, スコアリング手法に加え、最大エントロピー法 (ME 法) の3つの分類器の分類結果を用いて分類を行い、精度を調査する。

## 2. 分類器

本研究で組み合わせて用いた、SVM, ME 法, スコアリング手法の3つの分類器について以下で説明する。

### 2.1 SVM を用いた分類

SVM は、文書分類の分野で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。Pang らは様々な機械学習器を用いて文書分類を行い、SVM を用いた場合に最も高い精度を得ている。本研究でも、ツールとして SVM <sup>light</sup> \*<sup>1</sup> を使用し、Pang らと同様に SVM による分類を行う。

### 2.2 ME 法を用いた分類

Pang らは、2.1 節で述べた SVM による分類手法とともに、ME 法を用いた分類手法でも実験を行っている。文書の集合を  $D$ , 出力であるクラスの集合を  $C$  とするとき、文書  $d(\in D)$  でクラス  $c(\in (positive, negative))$  となる事象  $(c, d)$  の確率分布  $p(c, d)$  を ME 法により推定する。本研究では、実装に Amis\*<sup>2</sup> を用いる。パラメータ推定には先行研究とは異なり GIS アルゴリズムを用いる。

### 2.3 スコアリング手法を用いた分類

箆島らは、式 1 を使ったスコアリング手法を用いて分類を行っている。

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

$Score(w_i)$  は、素性  $w_i$  が肯定/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/否定的評判に偏って出現するほど、大きな正/負の値になる。文書中の素性のスコアの総和を求め、その値が正の値ならば肯定的文書、負の値ならば否定的文書としている。本研究では、この手法に対して、形容詞のスコアの重み変更と、 $\chi^2$  検定による素性の選別という2つの拡張を行っている [6]。

## 3. 複数の分類結果を組み合わせる手法

前節で述べた、SVM と ME 法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手法について、以下で説明する。

### 単純多数決

単純に、3つの分類器の多数決で最終的な結果を決める。

### 重み付き多数決

単純多数決では、個々の分類器の確からしさが全く考慮されていない。著者らはこれまでに、分類器の確からしさの指標として境界線からの距離を用いる手法を提案している。境界線からの距離に近いほど分類結果の信頼度が低いという仮定の基、分類を行っている。本研究では、重みに境界線からの距離を利用した重み付き多数決により分類を行う。重みは、それぞれの分類器で以下のものを用いる。

スコアリング 文書のスコアの絶対値

連絡先: 九州工業大学 情報工学部 知能情報工学科

〒 820-8502 飯塚市川津 680-4

Email: {k.tsutsumi, shimada, endo}@pluto.ai.kyutech.ac.jp

\*<sup>1</sup> <http://svmlight.joachims.org/>

\*<sup>2</sup> <http://www-tsujii.is.s.u-tokyo.ac.jp/amis>

表 1: 実験結果

分類方法	正解率
SVM	82.2%
スコアリング	83.4%
ME 法	80.5%
単純多数決	85.8%
重み付き多数決	86.4%
重み付き多数決 (手動)	87.5%
(先行研究)	
Matsumoto ら [4]	88.3%
Bai ら [5]	87.5%

SVM 分離超平面からの距離  $\times l$

ME 法  $|p(\text{positive}, d) - p(\text{negative}, d)| \times m$

$l, m$  は重みの正規化のために用いた定数である。訓練データによる予備実験から得られた境界線からの距離の平均値を基にして、自動的に値を決定する。

#### 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<sup>\*3</sup>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、Pang らと同様に、全レビュー中で出現回数4以上の単語を抽出し、否定文に対する前処理を行っている。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。個々の分類器 (SVM, スコアリング手法, ME 法), 多数決による分類手法 (単純多数決, 重み付き多数決, 重み付き多数決 (手動)) の計6つの手法による分類精度の調査を行った。重み付き多数決 (手動) は、交差検定の際に各 fold ごとの正解率が最大になるように定数  $l, m$  を手動で設定した場合の結果である。結果を表1に示す<sup>\*4</sup>。同じデータセットを用いて実験している研究で最も高い精度を得ているものに、部分依存木を素性に用いている Matsumoto ら [4], ペイジアンネットワークの1つである MBDAG を用いた Bai ら [5] の研究があり、それぞれの手法で最も精度が良い結果についても併記する<sup>\*5</sup>。

結果より、複数の分類器を用いた場合、単純な多数決でも個々の分類器を単体で用いた場合と比較して高い精度を得ることができた。加えて、重み付け多数決を行うことで単純多数決よりも若干精度が向上することを確認できた。しかし、 $l, m$  を正解率が最大になるように設定した場合の精度 87.5% と比較すると、1%以上の精度の開きがあり、最適な  $l, m$  を推定する手法の検討が必要であることが分かる。また、今回、重み正規化の際、定数倍するという単純な方法をとっているが、それぞれの分類器の出力値は全く違う意図をもって出力された値である。分類器の信頼度をより正確に反映させるためには、それぞれの値を適切に近似・正規化する手法についても議論が必要である。例えば、Platt ら [7] は SVM の出力を Sigmoid 関数を用いて確率値に近似することの有効性を実験的に示しており、このような手法を提案手法に適用することが考えられる。

<sup>\*3</sup> <http://www.cs.cornell.edu/people/pabo/movie-review-data/>

<sup>\*4</sup> Pang らの先行研究での精度は SVM が 82.9%, ME が 80.4% であり、本論文における追試結果とは異なっている。これは、素性選択の際の否定文の前処理などが完全に再現できなかったためだと考えられる。

<sup>\*5</sup> この論文での Matsumoto らの結果は各 fold ごとの最適なパラメータを設定した場合の精度で、提案手法の重み付き多数決 (手動) に相当する。

表 2: 正解率の上限

組み合わせる分類器	上限
スコアリング+SVM	91.6%
スコアリング+ME 法	91.8%
SVM+ME 法	92.2%
スコアリング+SVM + ME 法	94.9%

今回の実験の結果では先行研究の精度と比べて若干精度が劣っていた。そこで、提案手法の本質的な有効性について議論する。本手法では、どの分類器でも正しい結果が出力されていない場合、最終的に正しく分類することはできない。言い換えれば、どれか一つの分類器が正しい結果を出力していれば、提案手法の精度はさらに向上する可能性がある。すなわち、正解率の上限は以下の式2で求められる。

$$\text{正解率の上限} = \frac{1 \text{ つ以上の分類器で正しく分類している数}}{\text{全テストデータ数}} \times 100 \quad (2)$$

結果を表2に示す。本手法による最高の精度 87.5% に比べ、正解率の上限は 94.9% となっており大幅に高い。この結果より、重み正規化手法の検討により、大幅な精度向上の可能性がある事が分かる。また、重み正規化手法の検討以外でも、Matsumoto らが素性に利用した部分依存木などを提案手法に適用することで、さらに精度が向上すると考えられる。

#### 5. おわりに

複数の分類器を利用した本手法を用いることによって、単分類器しか用いない分類の場合よりも大幅に精度が向上する事を確認できた。今後は、組み合わせる分類手法や素性の追加、重み正規化手法の検討を行う。

#### 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] B. Pang, L. Lee, and S. Vaithyanathan. "Thumbs up? Sentiment Classification using Machine Learning Techniques". Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP) pp.79-86, 2002.
- [3] 箆島 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] Shotaro Matsumoto, 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.
- [5] Xue Bai, Rema Padman, and Edoardo Airoldi: Sentiment Extraction from Unstructured Text using Tabu Search-Enhanced Markov Blanket, Proceedings of the International Workshop on Mining for and from the Semantic Web (MSW2004), 2004.
- [6] 堤公孝, 嶋田和孝, 遠藤 勉: 分類の信頼度を利用した複数の分類器によるレビュー記事の自動分類, 人工知能基礎問題研究会, SIG-FPAI-A601, pp.27-32, 2006.
- [7] J. Platt: Probabilistic outputs for support vector machines and comparison to regularized likelihood methods, In Advances in Large Margin Classifiers, MIT Press, 1999.