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

2003－DBS－130 (21)
2003－ＦＩ－ 71 (21)
2003／5／23

SVM（Support Vector Machine）を用いた経済記事の著者の見解に基づく分類

中嶋 琢美　酒井 浩之　増山 繁

本研究では，著者の見解によるテキストの分類を目指す．今回は対象を経済記事とし，景気動向に対する見解別に分類を行った．分類は，最も単純な，景気が回復するであろうという見解と，景気が悪化するであろうという見解の 2 種類とした．景気が回復するであろうという見解の記事と，景気が悪化するであろうという見解の記事では，よく使用される語が異なるため，そこに着目した分類手法を考案した．本稿では，手法と，実験結果について報告をする．実験の結果，語の頻度を素性値としたSVM(Support Vector Machine) を用いた手法で，良好な結果が得られた．

A classification method based on the view of the author of each newspaper article on economics

Takumi Nakajima　Hiroyuki Sakai　Shigeru Masuyama

We propose a classification method of texts based on the view of the author. We treat newspaper articles on economics, and classify them into two classes, a class of articles where the author's view of each article is that business will recover, and a class of the author's view of each article that business will get worse, respectively.

豊橋技術科学大学 知識情報工学系
Department of Knowledge-based Information Engineering,
Toyohashi University Technology

1　はじめに

近年，莫大な量の機械可読なテキストが存在している．これらの情報を有効に活用するために，テキストマイニングの研究が活発に行われている [1]．テキストを分野毎に分類する研究 [2] は古くから行われてきたが，近年，同分野のテキストから著者の見解を分析する研究が行われるようになってきた．この研究の一環として，製品やサービスに対する消費者による評価文を対象としたテキストマイニングの研究がさかんに行われている [3] [4] [5]．

テキストからの主観的評価文の自動抽出に関する研究 [3] では，テキストから製品の評価文のみを抽出し，テキストを読む側の負担の軽減を目指している．この研究は，著者の見解に当たる部分を抽出するものであるが，見解を機械的に分類することは想定していない．主観的評価文からの意見抽出に関する研究 [4] では，製品やサービスに対する評価文の中でも，不満や否定的な意見を抽出することに特化している．また，製品に対する評価文が肯定的か否定的かを調べる試み [5] が行われているが，対象が製品への評価文のみで，人手で分類用の辞書を作成することを想定している．

上記の研究は，扱うテキストを消費者による評価文に限定したテキストマイニングであるが，本研究では，それ以外のテキストで著者の見解に基づく自動分類を目指す．大量のテキストが見解別に分類されていると，論文執筆の際の参考資料とする場合や，アナリストがテキストを基に分析を行う場合等に有用である．

この目標を実現させるための第一歩として，今回は対象とするテキストを経済記事に限定し，分類は，景気が回復するという見解 (これ以降 Positive と呼ぶ)，景気が悪化するという見解 (これ以降 Negative と呼ぶ) の最も単純な 2 種類とした．そして，対象とする記事は必ずどちらかに分類できるものに限定した．

2　手法考案にあたっての検討

手法考案にあたって，Positive，Negative な記事の特徴を調べた．その結果，それぞれに，多く出現する語があることが分かった．例えば Positive な記事では「回復」「成長」等の語が多く見られ，Negative な記事では「悪化」「減少」等の語が多く見られた．そこで，語の出現傾向によって記事を分類できると仮定した．あらかじめ学習用にいくつかのテキストを見解別に分類しておき，それぞれに多く出現する語を調べれば，その語の頻度を基に未知のテキストも分類できるはずである．したがって今回の場合は，あらかじめ Positive，Negative のいずれかに分類されている記事で，それぞれに多く出現する語を調べれば，その語の頻度を基に未知の記事も分類できるはずである．

ここで考えられる問題点として，Positive な記事で「赤字の減少」や,「失業者の減少」等のように，本来 Negative な記事で多く出現する語が使われる可能性がある．しかし，そのような文が記事中に出現したとしても，他の文では,「回復」,「成長」等の語が使用されるため，その影響は小さく，問題なく分類できるものと考えた．

また，接尾辞やサ変動詞等，品詞によっては，分類に役立たないものがあるため，分類に利用する語を品詞によって選択することを考えた．

3　予備実験

2 節で述べた仮定の有効性を調べるために予備実験を行った．実験は語の頻度の合計数によって分類する方法と，語の頻度を素性として SVM によって分類する方法の 2 種類について行った．以下に手法を示す．

3.1　頻度の合計値による分類手法

Step 1. 学習用とテスト用の記事 (これ以降，それぞれ，学習用記事，テスト用記事と呼ぶ) を人手で用意する．

Step 2. 集めた記事を形態素解析する．

Step 3. 分類に使用する語 (これ以降分類語と呼ぶ) を学習用記事から自動的に抽出する.

Step 4. テスト用記事における分類語の頻度の合計値で分類する. □

分類語抽出と頻度の合計値による分類について以下に詳しく述べる.

- 分類語抽出
  分類語は, 学習用記事を形態素解析した結果から抽出する. 品詞の種類を基に, 全学習用記事から分類語の候補 (これ以降分類語候補と呼ぶ) を選択する. ここで, 学習用記事の Positive な記事で一定回数以上出現するという条件 (これ以降分類語条件と呼ぶ) を満たした形態素を Positive 分類語, Negative な記事で分類語条件を満たした形態素を Negative 分類語と呼ぶことにする. なお, Positive, Negative な記事の双方で多く出現する分類語は稀であったため, そのような語に対して特別な措置はとらないことにした.

- 頻度の合計値による分類
  テスト用記事中の Positive 分類語, Negative 分類語の頻度をそれぞれ加算していく. 最終的に Positive 分類語の頻度が Negative 分類語の頻度を上回った場合は Positive であると分類し, その逆であれば Negative であると分類する.

3.2 頻度を素性としたSVMによる分類手法

Step 1. 学習用記事, テスト用記事を人手で用意する.

Step 2. 集めた記事を形態素解析する.

Step 3. 分類語を学習用記事から自動的に抽出する.

Step 4. 学習用記事中の分類語の頻度を基に SVM によって学習する.

Step 5. テスト用記事中の分類語の頻度を基に SVM で分類する. □

SVM による学習, 分類は, TinySVM[1]を使用して Positive 分類語と Negative 分類語の頻度を素性値とした.

3.3 予備実験分類結果

91,92,94,96 年の日経新聞から, 学習用記事は Positive, Negative な記事を 15 記事ずつ, テスト用記事は 30 記事ずつ集めて実験を行った. 形態素解析器は JUMAN[2] を使用した.

今回は分類に有用であると考えられる語がサ変名詞に多かったため, 分類語候補をサ変名詞に限定した. 分類語条件を変化させた場合について, 表 1, 表 2 に結果を記す.

表 1: 頻度の合計値による分類実験結果

分類語条件	分類語数	Positive 正解数	Negative 正解数
5 回以上	37	27	16
10 回以上	18	26	25
15 回以上	9	27	22

表 2: SVM による分類実験結果

分類語条件	分類語数	Positive 正解数	Negative 正解数
5 回以上	37	26	26
10 回以上	18	28	22
15 回以上	9	24	30

[1]奈良先端大の工藤らが開発.
[2]http://www-nagao.kuee.kyoto-u.ac.jp/

3.4 予備実験に対する考察

どちらの手法も，分類語条件によって多少の正解数の差が見られるものの，結果は良好であり，仮定は正しかったといえる．特にSVMによる分類手法の方は，分類語条件がいずれの場合も高い正解数となった．また，頻度の合計値による分類手法では，Positiveな語とNegativeな語の頻度の合計値が同数となり，分類できない場合が見られた．それに対し，SVMによる分類手法ではそのような場合は発生しなかった．このような点から，SVMによる分類手法の方が優れていると考え，この手法でさらなる実験を行った．

4 本実験

予備実験では，分類語候補をサ変名詞に限定していたが，これ以外の品詞の語も分類語候補に加えることでさらに正解数を増やすことができるのではないかと考えた．そこで，分類語候補としてサ変名詞と以下に記す品詞を付け加えて実験を行った．

- カタカナ語
  カタカナ語の中には「プラス」や「マイナス」等分類に役立ちそうな語が含まれていると考えた．
- 複合語
  「後退」と「景気後退」という語を比べた場合，後者の方がよりNegativeな状態を表現している．そこで複合語が分類に役立つのではないかと考えた．本研究における複合語の定義を以下に詳しく記す．

複合語の定義

記事を形態素解析したものを出現順に全て取り出す．取り出した形態素の列をm_1，m_2，m_3，...，m_nとする．この時$m_i(1 \leq i < n)$が普通名詞で，かつm_{i+1}がサ変名詞であった場合，この2つの形態素を結合することによって作られるものを複合語とする．

4.1 本実験分類結果

学習用記事，テスト用記事共に予備実験と同じ記事を使用して実験を行った．分類語条件を変化させた場合について，表3，表4に結果を記す．

4.2 本実験に対する考察

サ変名詞以外の語を分類語候補に加えた場合，若干の実験結果の変化が見られたが，大幅に結果が改善されることはなかった．

カタカナ語の場合は「プラス」と「マイナス」以外の語で，分類に役立つような語がなかったため，結果に変化がみられなかったものと考えられる．

分類語候補が複合語の場合，「景気回復」，「景気後退」など分類に役立ちそうな語も存在したが，「個人消費」，「中間決算」などPositive，Negativeに関係のない語も多くみられた．これが原因となり結果に変化がみられなかったものと考えられる．

5 考察

実験結果が良好であったことから，語の頻度を素性値とした，SVMを用いた分類は有効であるといえる．特に，分類語候補がサ変名詞のみで分類語条件が15回以上の場合は，Negativeな記事を全て正しく分類することに成功している．

結果が良好であった理由としては，記事は全て同じ新聞社の記事から集めたものであったため，記事ごとの表記のゆれが少なく，使用された語も似ており，文の長さにも大きな差がなかったことが挙げられる．このような点から，本手法は，対象となるテキストが上記の点で似ている場合には有効であるといえるが，この条件を満たさなかった場合の精度の保証はない．また，

表 3: サ変名詞とカタカナ語実験結果

サ変名詞 分類語条件	カタカナ語 分類語条件	分類語数	Positive 正解数	Negative 正解数
5 回以上	5 回以上	41	27	22
10 回以上	10 回以上	20	28	22
15 回以上	15 回以上	11	28	27

表 4: サ変名詞と複合語語実験結果

サ変名詞 分類語条件	複合語 分類語条件	分類語数	Positive 正解数	Negative 正解数
5 回以上	2 回以上	64	26	26
5 回以上	3 回以上	51	26	26
5 回以上	4 回以上	45	25	26
10 回以上	2 回以上	45	27	22
10 回以上	3 回以上	32	27	23
10 回以上	4 回以上	26	27	25
15 回以上	2 回以上	36	29	22
15 回以上	3 回以上	23	29	25
15 回以上	4 回以上	17	28	25

新聞記事ということで，誤字，脱字が発生する確率が無視できるほど小さなものであったことが挙げられる．

今後は，他社の新聞社の記事を混ぜた場合の精度の実験や，経済記事以外で本手法を適用した場合の実験が望まれる．

6 まとめ

本稿では，教師あり学習による，著者の見解別にテキストを分類する手法について述べた．この手法は，分類に使用する語を学習用記事から自動的に抽出という特徴を持つ．景気動向に関する経済記事を対象とした実験で，語の頻度を素性値としたSVMを利用した分類で良好な結果が得られた．今後の課題として，対象とするテキストを変更した場合の実験を行うことが挙げられる．

謝辞

言語データとして，日本経済新聞 CD-ROM 版の使用を許可して頂いた日本経済新聞社に感謝いたします．

参考文献

[1] 市村 由実，長谷川 隆明，渡部 勇，佐藤 光弘: テキストマイニング－事例紹介，人工知能学会誌，16 巻，2 号，pp.192-200(2001)．

[2] 平瀬順，春野雅彦: Support Vector Machine によるテキスト分類における属性選択，情報処理学会論文誌，Vol 41，No. 4，pp. 1113-1123(2000)．

[3] 村野 誠治，佐藤 理史: 文型パターンを用いた主観的評価文の自動抽出，言語処

理学会第 9 回年次大会発表論文集，pp.67-70(2003)．

[4] 舘野 昌一:「お客様の声」に含まれるテキスト感性表現の抽出方法，言語処理学会第 9 回年次大会発表論文集，pp．71-72(2003)．

[5] 長江 朋，望月 源，白井 清昭，島津 明: 製品コンセプトと製品評価文章の関係の分析，言語処理学会第 8 回年次大会発表論文集，pp．583-586(2002)．

付録

分類語を以下に記す．

- 分類語候補:サ変名詞
  分類語条件:5 回以上

  悪化 回答 回復 改善 期待 経営 決算 減少 雇用 後退 好転 実施 受注 修正 所得 消費 上昇 成長 生産 設備 総合 増加 調査 調整 低下 低迷 投資 同期 発表 判断 販売 不足 分析 輸送 予想 予測 連続

- 分類語候補:サ変名詞
  分類語条件:10 回以上

  悪化 回答 回復 改善 減少 後退 実施 消費 上昇 成長 生産 調査 低迷 投資 同期 判断 予想 予測

- 分類語候補:サ変名詞
  分類語条件:15 回以上

  悪化 回復 改善 減少 上昇 生産 調査 判断 予測

- 分類語候補:サ変名詞
  分類語条件:5 回以上
  分類語候補:カタカナ語
  分類語条件:5 回以上

  悪化 回答 回復 改善 期待 経営 決算 減少 雇用 後退 好転 実施 受注 修正 所得 消費 上昇 成長 生産 設備 総合 増加 調査 調整 低下 低迷 投資 同期 発表判断 販売 不足 分析 輸送 予想 予測 連続 プラス ベース ポイント マイナス

- 分類語候補:サ変名詞
  分類語条件:5 回以上
  分類語候補:複合語
  分類語条件:2 回以上

  悪化 回答 回復 改善 期待 経営 決算 減少 雇用 後退 好転 実施 受注 修正 所得 消費 上昇 成長 生産 設備 総合 増加 調査 調整 低下 低迷 投資 同期 発表 判断 販売 不足 分析 輸送 予想 予測 連続 下方修正 期低迷 期予想 季節調整 業種別 景気悪化 景気回復 景気後退 景気判断 景気予測 景況調査 景況判断 経済観測 経済経営 建設業保証 県民所得 個人消費 工業生産 国内販売 在庫調整 住宅着工 小企業総合 小幅改善 中間決算 動向調査 能力増強 本社調査

- 分類語候補:サ変名詞
  分類語条件:5 回以上
  分類語候補:複合語
  分類語条件:4 回以上

  悪化 回答 回復 改善 期待 経営 決算 減少 雇用 後退 好転 実施 受注 修正 所得 消費 上昇 成長 生産 設備 総合 増加 調査 調整 低下 低迷 投資 同期 発表 判断 販売 不足 分析 輸送 予想 予測 連続 下方修正 景気後退景況調査 景況判断 経済観測 建設業保証 個人消費 中間決算