SVMを用いた株価短報における意見文と事実文の抽出

嶋田 康平[1]　　岡田 真[2]　　橋本 喜代太[3]
大阪府立大学
{[1] st301020@edu, [2] okada@mi.s, [3] hash@kis}.osakafu-u.ac.jp

1. はじめに

近年，電子商取引が盛んになっている．それに伴い，株取引に関してもウェブ上で情報を公開されており，株価にかかわるような記事を配信するサービスが日常的に用いられるようになってきている．

株価に関する研究もなされており，廣川ら[1]は新聞記事を業種別に分類することで，業種によって単語が株価動向に与える影響の違いや新聞記事と株価動向の関係性のついての解析をおこなっている．

株価にかかわるような記事の1つに，株価短報と呼ばれるものがある．これは株価にかかわるさまざまなニュースを希望者に配信するものであり，読者はその内容を株取引の参考にする．

株価短報中のニュース記事には事実を伝える事実文とそれを基にした記事記述者の推測や予測を表す意見文が含まれており，これらを区別して提示することは，読者にとって有益である．記事の読者がそれらを混同して株取引の判断を誤ることを防ぎ，予測と異なる展開による不利益を避けることができる．また，ある1つの事実に対して複数の意見が存在する場合に，短報中からそれらを抽出，整理し，利用者に提示することで，それらの比較検討することにより，多角的かつ状況に適した判断を行うことが可能となる．

そこで本研究では株価短報中の事実と意見を区別して抽出，整理するための前段階として，事実文と意見文の自動分類手法について提案する．今回は機械学習手法の 1 つであるサポートベクトルマシン(SVM)を用いて，事実文と意見文の自動分類を行う．

以下，2 章では株価短報とそれに含まれる意見文と事実文について，3 章では本研究で自動分類に用いたサポートベクトルマシン(SVM)の概要について説明し，4 章で実験と考察について述べる．5 章でまとめと今後の課題について説明する．

2. 株価短報と意見文と事実文

株価短報は株式取引を行う人々に送付されるテキストベースのニュース記事である．本研究で用いた記事では，そこに含まれるニュースの長さは短いもので 20 字前後，長いものでは 1,000 字以上のものまであり，1 つのニュースの平均文字数は 803 字であった．

意見文と事実文

株価短報に対するニュース記事には株価の事実をそのまま述べてある事実文と，それら事実に対する記事著者の意見もしくは予想が述べてある意見文が含まれている．

本研究では前処理として，対象となる文について 1 文ずつ人手で意見文と事実文のラベル付けを行った．ニュース記事 1000 記事に含まれる 7440 文についてラベル付けを行い，その結果を表 1 に示す．記事は 100 記事ごとに分けてデータを記述してある．7440 文中の内訳は，意見文は 847 文，事実文は 6593 文となった．

ラベル付けの基準

ラベル付けでは，あらかじめ意見文と判定するための基準を設けておき，それに当てはまらない文は事実文と判定した．

意見文と判定する基準として，本研究では文末表現に着目し，文末に以下の表現が含まれている場合

に意見文として判定した．図 1 に判定基準である文末表現を示す．

表 1　意見文と事実文

記事	意見文	事実文
①	93	630
②	97	541
③	96	804
④	98	713
⑤	36	596
⑥	112	823
⑦	25	377
⑧	127	890
⑨	70	403
⑩	93	816
合計	847	6593

・～もよう。　・～よう（だ）。
・～しておきたい。　・～そう（だ）。
・～かもしれない。　・～とみられる。
・～だろう。　・～したい。

図 1　意見文に含まれる文末表現

3. SVM

今回は機械学習を用いて分類を行う．本研究ではサポートベクトルマシン（SVM）を使用する．SVM は近年の自然言語処理関連の研究でも盛んに使われている．

SVM は Vapnik[2]によって提案されたデータを 2 つのクラスに分類する教師あり学習アルゴリズムである．図 2 にサポートベクトルマシンの概念図を示す．正例と負例の含まれたベクトルの学習データが与えられ，それらの正例と負例を区切る超平面を計算し，その超平面によって未知データのクラスを推測する．SVM は高次元のベクトル空間であっても超平面で分類することができ，高い汎化能力を持つことが知られている．自然言語処理のベクトル化は高次元になることが多いため，自然言語処理の研究において SVM が用いられることは多い．本研究では，カーネル関数と組み合わせて SVM を利用する．カーネル関数は，特徴空間における内積をデータの座標の明示的な計算を経由せずに，データから直接計算する手段を与える．カーネル関数を用いることで，内積を計算する際の計算量を少なくできる．

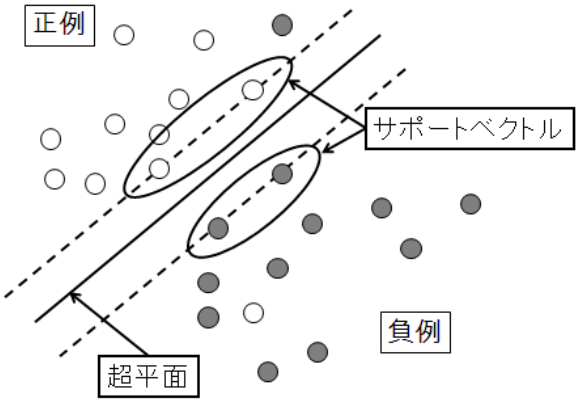

図 2　サポートベクトルマシンの概念図

4. 実験と考察

4.1　実験

株価短報の記事 1000 記事(7440 文)に対して人手で意見文と事実文のラベル付けを行い，それらを学習用データおよびテストデータとして用い SVM により分類を行い，どの程度の精度で分類できるか実験により調査した．今回の実験ではライブラリとして LIBSVM, Kernel 関数は Gaussian(RBF) Kernel を使用した．

今回の実験では，全 1000 記事を 10 分割し，10 分割交差検証を行うことにした．全 1000 記事のうち 900 記事を学習データ，残りの 100 記事をテストデータとして，それぞれ形態素解析エンジン MeCab を用いて形態素に分割し，学習データの形態素群を基にサポートベクトルを生成し，それを用いてテストデータの分類を行った．この時，図 1 に示した文末表現を一つの要素として認識させていない．

今回は 10 回検証を行い，意見文，事実文共に分類の精度をそれぞれ求めた．表 2 に実験結果を示す．

表 2 の番号は，表 1 の 100 記事の番号に対応し，交差検証の各試行ではその番号のデータをテストデータとして用いた．

表 2　実験結果（精度）

記事	意見文		事実文	
	正例/総数	精度(%)	正例/総数	精度(%)
①	50/ 93	53.8	623/630	98.9
②	56/ 97	57.7	541/541	100
③	62/ 96	64.6	804/804	100
④	66/ 98	67.4	713/713	100
⑤	12/ 36	33.3	596/596	100
⑥	67/112	59.8	820/823	99.6
⑦	9/ 25	36.0	377/377	100
⑧	72/127	56.7	890/890	100
⑨	48/ 70	68.6	401/403	99.5
⑩	2/ 93	2.2	814/816	99.8
平均	444/847	52.4	6579/6593	99.8

4.2　考察

意見文の全体の精度の平均値 52.4%となった．ここで，⑤と⑦と⑩の意見文の分類精度がそれぞれ 33.3%，36.0%，2.2%と低く，この 3 つの結果が全体の平均精度を引き下げていることが実験結果から読み取れる．このうち⑤と⑦については，テストデータ中の意見文数がそれぞれ 36 文と 25 文と他のデータの 3 分の 1 程度と少なく，さらに，事実文に対する意見文の割合もそれぞれ 6%と 6.6%と全体の平均値の 12.8%よりも低い．これら 3 つの結果を除いた精度の平均値は 61.2%ではあるが，これでも十分な精度とはならず，意見文推定の精度向上の余地が大きく残る結果となった．

事実文については，全体の分類精度の平均値が 99.8%となっているが，もともとのデータにおける事実文の占める割合が約 88.6%と圧倒的多数であったことが強く影響していると考えられる．

これに対して意見文は全体として精度が低いが，これは単語一つずつを特徴ベクトルとするスパースな空間の中で正解数が 1 割程度と低いようなデータの場合には，一般に精度が大きく落ちることが知られている．このため，このようなデータでは前処理として特徴空間の次元圧縮を行うなどが必要されると考えられる．

さらに今回はベクトル生成時に図 1 に示した特徴的な文末表現を一つの要素として認識させておらず，単純に形態素に分解して学習を行っている点も悪影響を与えていると考えられる．

5. おわりに

本稿では，株価短報の記事について，対象となる文に対し意見文と事実文に人手でラベル付けをおこない，機械学習手法 SVM のための学習データと実験データを作成した．そして，意見文と事実文の分類の精度を実験により求め，考察をおこなった．

今後の課題としては，分類精度のさらなる向上が挙げられる．例えば，品詞や文末表現に着目し，それらの与える影響の検証，SVM の Kernel 関数やパラメータを変えた場合の精度の検証，また，川口ら[3]がおこなった手法を参考に，学習データの意見文と事実文の数の統一した場合の精度の検証，さらに学習データの総数を増やして検証を行うことなどが挙げられる．そのほかには，データ量の増加に伴う計算時間の増加への対策として，より効率の良いアルゴリズムを考慮する必要もある．

参考文献

[1]廣川敬真,吉田稔,山田剛一,増田英孝,中川裕志 ：“業種別による新聞記事と株価動向の関係の解析”, 言語処理学会 第 16 回年次大会， pp.1070-1073, (2010)

[2]V. N. Vapnik , The Nature of Statistical Learning Theory, 2nd ed., Springer-Verlag, New York, (2000)

[3]川口敏弘,松井藤五郎,大和田勇人 ：“SVM と新聞記事を用いた Weblog からの意見文抽出”,第 20 回人工知能学会全国大会, 1A3-3, 4 ページ, (2006)