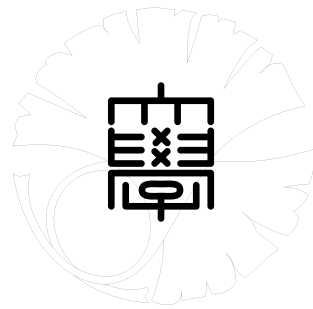


修士論文

為替ニュース記事を用いた
SVMによる株価動向予測

Using news articles on foreign exchange
to predict trend of stock prices by SVMs



2014年2月6日

指導教員 伊庭 齊志 教授
ダヌシカ ボレガラ 講師

東京大学大学院 情報理工学系研究科

電子情報学専攻
48-126404

石黒 祐輔
Yusuke Ishiguro

Abstract

There are many research about predicting stock prices using text mining, but there is no significant research. In many research, they use all kinds of news as text sources, and there is no work using specific kind of news. I supposed that using all kinds of news is not good for predicting stock prices. In this research, I proposed methods to predict the trend of stock prices by SVMs using news articles on foreign exchange broadcast on the web.

I proposed four methods with the system. First of all, I use the news broadcast by stock company which used by investors, restricted to regular news of foreign exchange rate. This is because the investors are the ones who move stock prices and the news they use are not popular newspaper, they use special news. Secondly, I use the title of the news not article, to convert to feature vector by Bag-of-Words. I supposed that there are many noises in articles and the title would be express contents more brief. Thirdly, I filtered some parts of speech based on morphological analysis program 'MeCab' with IPA dictionary. This is for removing stop words from feature values by part of speech, not term itself. At last, I set lower bound and upper bound as thresholds to frequency of terms for filtering feature values, and whenever the frequency of term is out of bound, it will be removed from feature values. This is a language independent method for removing stop words and rare words.

I did some experiments changing the combination and parameters of proposed methods to predict the trend of issues which chosen for the Nikkei 225 index. As the result, I confirmed efficacy of all proposed methods. For the threshold of frequency of terms, it was best to set 0.06 for the lower bound and 1.0 for the upper bound.

But last of all, the premise of this research, "The export and import companies are affected by foreign exchange rates, domestic demand companies are not." was partly negated. It means that there is a need for future research with the relationship between foreign exchange rate and individual issues of stock.

目次

第 1 章	序論	1
1.1	研究背景	2
1.2	本研究の目的	3
1.3	本論文の構成	4
第 2 章	株式市場	6
2.1	市場原理と投資家	7
2.2	投資判断と株価への影響	8
2.3	日本の株式市場と指標	8
2.4	外貨為替相場と株価の関係	10
第 3 章	テキストマイニング	12
3.1	データマイニング	13
3.2	データマイニングにおける解析手法	14
3.3	サポートベクターマシン	15
3.4	テキストマイニング特有の解析手法	16
3.5	日本語固有の問題	18
第 4 章	提案手法	20
4.1	既存研究の問題点	21
4.1.1	新聞媒体の問題	21
4.1.2	制限のないニュースの対象化	22
4.1.3	確立したテキストマイニング手法の非存在	22
4.2	提案手法の全体像とアルゴリズム	23
4.3	自然言語処理	24
4.3.1	単語のフィルタリング	24
4.3.2	入力テキスト	25

4.4	教師信号	26
4.4.1	分類クラス	26
4.4.2	株価動向の定義	26
4.4.3	気配値	28
4.5	SVMによる学習	28
4.5.1	学習データのスケールリング	28
4.5.2	SVMライブラリとパラメータ	29
4.5.3	n 分割交差検定	30
4.6	評価方法	30
4.6.1	正答率	31
4.6.2	平均F値	31
4.6.3	仮想トレード	32
第5章	株価動向予測実験	33
5.1	データ	34
5.1.1	ニュースデータ	34
5.1.2	株価データ	34
5.1.3	データ期間	34
5.2	実験方法	35
5.3	ベースライン	36
5.3.1	ランダム予測	36
5.3.2	Bag-of-Wordsのみで予測	36
5.4	提案手法の個別適用実験	38
5.4.1	タイトルのみで予測	38
5.4.2	品詞フィルタのみで予測	39
5.4.3	単語出現頻度フィルタのみで予測	39
5.5	提案手法の複数種組み合わせ実験	42
5.5.1	タイトル+品詞フィルタで予測	42
5.5.2	品詞フィルタ+単語出現頻度フィルタで予測	42
5.5.3	タイトル+単語出現頻度フィルタで予測	43
5.6	全手法の適用実験	43
5.6.1	単語出現頻度のパラメータ実験	43

5.6.2	予測期間のパラメータ実験	44
5.7	最優秀結果の詳細	47
5.8	日経平均 223 銘柄での実験	48
第 6 章	考察	50
6.1	為替ニュース	51
6.2	タイトルの利用	51
6.3	品詞フィルタ	52
6.4	単語出現頻度フィルタ	52
6.5	平均 F 値	53
6.6	予測期間	53
第 7 章	結論	55
7.1	まとめ	56
7.2	今後の展望	57
	参考文献	59
	発表文献	63
	付録	64
A	MeCab における IPA 辞書の品詞体系	65
B	F 値の求め方	68
C	日経平均 225 銘柄	69

目次

1.1	効率的市場仮説における理想的な状態の価格推移	3
2.1	投資家の行動原理の一例	7
2.2	日経平均と TOPIX の過去数年の推移	9
2.3	TOPIX と米ドル/円の推移 [1]	11
2.4	TOPIX と米ドル/円の相関性 [1]	11
3.1	SVM による 2 次元空間の分離平面の例	15
3.2	SVM による様々な分離平面	16
3.3	SVM によって線形分離できない例	16
3.4	カーネル関数によって線形分離できる空間へ写像	17
4.1	提案手法の全体図	23
4.2	株価動向の求め方のサンプル	27
4.3	n 分割交差検定	30
5.1	2013 年の日経平均の推移	36
5.2	ランダム予測で予測期間を変化させた実験結果	37
5.3	BoW のみで予測期間を変化させた実験結果	37
5.4	BoW+タイトルで予測期間を変化させた実験結果	38
5.5	BoW+品詞フィルタで予測期間を変化させた実験結果	39
5.6	BoW+単語出現頻度で出現頻度のしきい値を変化させた実験の訓練セット正答率	40
5.7	BoW+単語出現頻度で出現頻度のしきい値を変化させた実験のテストセット正答率	40
5.8	BoW+単語出現頻度で出現頻度のしきい値を変化させた実験の平均 F 値	41
5.9	BoW+タイトル+品詞フィルタで予測期間を変化させた実験結果	42

5.10 BoW+品詞フィルタ+単語出現頻度フィルタ [0.05, 1.0] で予測期間を変化させた実験結果	43
5.11 BoW+タイトル+単語出現頻度フィルタ [0.05, 1.0] で予測期間を変化させた実験結果	44
5.12 BoW+タイトル+品詞フィルタ+単語出現頻度で出現頻度のしきい値を変化させた実験の訓練セット正答率	45
5.13 BoW+タイトル+品詞フィルタ+単語出現頻度で出現頻度のしきい値を変化させた実験のテストセット正答率	45
5.14 BoW+タイトル+品詞フィルタ+単語出現頻度で出現頻度のしきい値を変化させた実験の平均 F 値	46
5.15 BoW+タイトル+品詞フィルタ+出現頻度フィルタ [0.06, 1.0] で予測期間を変化させた実験結果	46
5.16 BoW+タイトル+品詞フィルタ+出現頻度フィルタ [0.06, 1.0] で予測期間を変化させた実験結果 (日経平均 223 銘柄)	49

表目次

2.1	為替と株価の関係	10
4.1	本研究でフィルタリングする品詞	25
4.2	2クラス分類で得られた結果の例	31
5.1	実験に用いた銘柄一覧	35
5.2	予測期間=30分, 単語出現頻度フィルタの下限值=0.06, 上限値=1.0の時の銘柄別の評価値	47
A.1	MeCabにおけるIPA辞書の品詞体系	65
A.1	MeCabにおけるIPA辞書の品詞体系	66
A.1	MeCabにおけるIPA辞書の品詞体系	67
B.1	混同行列	68
C.1	日経平均225銘柄(2012年10月30日時点)	69
C.1	日経平均225銘柄(2012年10月30日時点)	70
C.1	日経平均225銘柄(2012年10月30日時点)	71
C.1	日経平均225銘柄(2012年10月30日時点)	72

第1章

序論

1.1 研究背景

市場経済が世界に導入されて長い時間が経ち、人類の経済活動の活発化に伴い金融市場も成長した。現在では金融市場の規模は非常に大きなものとなり、それが人類の経済活動に与える影響が絶大なものとなった。そのため金融市場のメカニズムに関する様々な研究が進められているが、金融市場は非常に複雑な事象の集合であり、その全容を解明するに至っていない。

そこで市場のモデル化に関する研究とは別に、何らかの手法を用いて市場の予測をすることによって市場と市場外で起こる事象の関係を結びつけるパラメータを特定し、金融市場のメカニズムの解明に役立てる研究が様々な側面から行われている。その手法の1つが、テキストマイニングである。

近年高度な情報社会化が進み、大量の情報がウェブ上に蓄えられるようになった。その中には情報として価値の有るものから無いものまで様々であり、そこから必要な情報を抽出することは難しい。データマイニングによって大量の情報を適切に処理できれば、そこから新しい知識を獲得する可能性もあり、既存の研究に役立てることができる。そこでウェブ上のデータから適切に情報を抽出するために、テキストマイニングに関する研究が様々な分野で行われおり、それは経済の分野でも例外ではない。特にウェブ上の情報をテキストマイニングすることによって市場の予測をすることができれば、市場のモデル化に貢献するからである。

ウェブ上から取得できる経済関連の情報としては様々なものがあるが、特に可能性があるのがニュースである。[2] それは、近代の金融理論の基礎を成している効率的市場仮説 (Efficient Market Hypothesis; EMH)[3] と呼ばれる、市場が過去から将来までのあらゆる情報を織り込んで動いているという仮説によって説明でき、この仮説は広く信じられているものとなっている。このため図 1.1 のように、新しい情報が来ない限り価格は一定であり、一度新しい情報が入って来たらすぐに新しい情報の価値を含んだ価格に遷移することを意味する。例えば“A社は今期は予想通り赤字”というニュースが市場に流れた場合でも、既に株価へ“今期赤字”という情報が織り込まれているため株価への影響は少ない。このような市場に“予期していない新しい情報”が舞い込めば、市場はすぐさま大きく反応する。この“予期していない新しい情報”を与えるのがニュースであるからである。

ウェブ上の情報をテキストマイニングし、市場を予測するという研究は比較的新しい分野であり、手法としては確立されていない部分が多い。本稿ではニュースを対象とするが、ニュース以外の研究も多くなされている。金融市場の種類も株式市場、外貨為替市場、債券

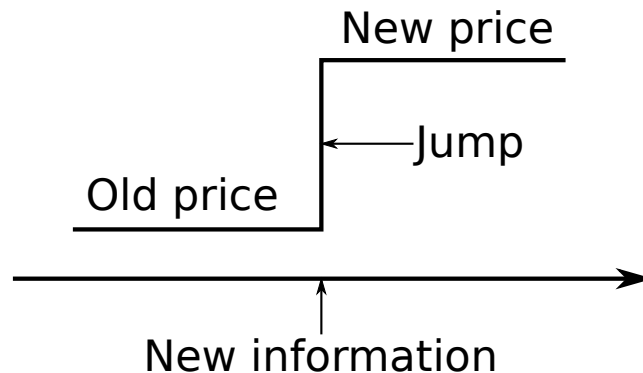


図 1.1: 効率的市場仮説における理想的な状態の価格推移

市場など数多くあるが、株式市場や外貨為替市場以外はあまり研究されていない側面がある。また国によっては必ずしもウェブや金融市場の性質が一致しない。そこで本研究では、日本語のニュースをテキストマイニングし、日本の株式市場、特に東証一部を対象とした市場予測を行うことにする。

1.2 本研究の目的

ニュースをテキストマイニングし、何らかの株価を正確に予測できるようになることが一番望ましいが、現実的には株価が定まるまでにニュース以外の様々な要因があるため、ニュースから得られる情報だけで株価を正確に予測することは難しい。そこでニュースの影響によって株価が上昇するか下降するかといった株価動向を予測するという単純化を図ることにする。

また世の中には未来のことから過去の事まで様々なニュースがあるため、あらゆるニュースをテキストマイニングの対象とすると予測モデルとしては非常に複雑で難しいものとなる。そこで外貨為替市場に関する市況のみを対象とすることによって、外貨為替市場と株式市場の関係に焦点を当てる。

以上より、本研究では外貨為替市場に関する市況のみをテキストマイニングし、機械学習の手法の一つであるサポートベクターマシン (SVM) によって様々な株価の動向を予測し、その正答率を評価する。

特に既存研究ではあまり着目されていなかった、ニュースの絞り込みを行うことや、あえて記事本文ではなくタイトルのみを用いること、また新たな評価尺度を用いることを本研究の貢献とする。

1.3 本論文の構成

本稿は以下のような構成である。

第2章

本研究の前提となるが情報処理の世界では馴染みの薄い金融市場および株式市場に関する知識を述べる。まず市場原理下における投資家の一般的な行動原理に対して、それに対して価格がどう反応するかを述べる。そしてその後日本の場合の株式市場の例として、東証とそれに関連した指標に関して説明する。最後に市場としては異なる、為替市場と株式市場の関係について述べる。

第3章

一般的なテキストマイニングのアプローチから、データマイニングする上で欠かすことのできない機械学習として、サポートベクターマシンについて説明する。この章では、基本的に言語非依存な手法として説明するが、最後に言語依存な問題として、日本語固有の問題を述べる。

第4章

本研究が提案する手法の詳細について説明する。まず手法の全体像から俯瞰した後、最初のステップから順を追って説明する。章の最後に、本研究の実験における評価手法について説明する。

第5章

実験結果について述べる。実験では、まず使用するデータについて説明した後、実験手法について説明する。そしてベースラインとして2種類の手法による評価を行う。その後提案手法を様々なパターンで組み合わせ、そのパターンに対して実験を行い、結果を評価することで提案手法の部分部分を個別に評価する。最後に提案手法全てを組み合わせた状態で実験を行い、その結果の詳細について考察する。

第6章

実験結果に対する考察を行い、提案手法に利点や欠点、そして今後の改良点について述べる。

第7章

本研究のまとめを述べ、今後の展望を述べる。

第2章

株式市場

まず情報系の世界では馴染みの薄い株式市場について、基礎的な予備知識を述べることにする。その後、本研究で重要となる為替市場と株価の関係について説明する。

2.1 市場原理と投資家

現代の市場経済においては、ある財(株式市場で言えば株式)が取引される場合、その価格と数量は基本的に需要と供給によって決定される。このように市場が自らを自動的に最適化¹するため、市場参加者は自らの利益のみを追求することで、結果的に市場全体がバランスを取れるように機能されている。

そのため金融市場においては、投資家らは利益最大化のために通常図2.1のように振る舞うと考えられる。まず投資家は、現在得られる目的の銘柄についての情報を集め、それらの情報に基づき売買、もしくは何もしないという行動を起こす。またニュースなどより新しい情報が入ってきた場合も、その情報を解釈し、それまでの情報と比較することによって行動を選択する。投資家が人間である限り²、その行動には理念・感情・思惑など様々な精神的要素が反映される。このため同じ情報に対して、投資家によって情報の解釈に差が発生し、また精神面で異なるので、その行動は一様なものとはならない。³

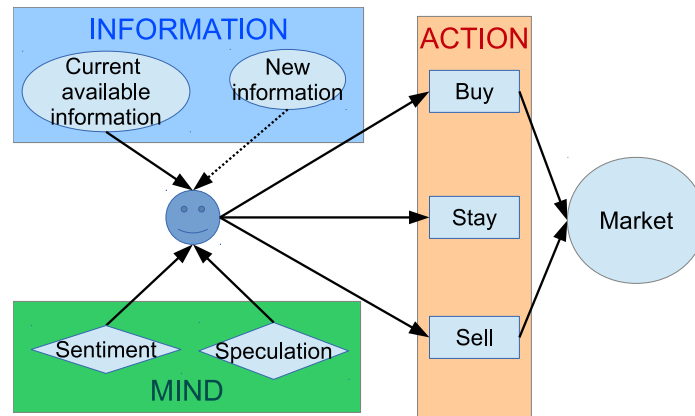


図 2.1: 投資家の行動原理の一例

例えば“A社赤字に転落”というニュースが市場に流れたとき、多くの投資家らがA社の株の売却を行う一方で、株価が下るのを狙って戦略的に株を買う投資家らもいる。このように同じ情報でも人によって選択する行動は異なるものとなる。

¹ただし過度な変化に対しては最適化が間に合わないこともあり、その場合全体の不利益に繋がるので、政府などが補助的に最適化を促す場合もある。

²機械がアルゴリズムに基づいて取引する場合も存在する。

³簡単な例であり、実際にはこれら以外にも様々な要素が関係する。

2.2 投資判断と株価への影響

前節では、新しい情報が入ってきた時の投資家らの反応について説明した。しかしながら実際に投資家が取引する場合には、新しい情報によって動かされる場合は少なく、現在の株価と投資家が考える将来の株価の差によって取引を行う。将来の株価を想定するためにそれまでに得られた情報を多角的に分析する。この情報を分析する手法としては、主に以下の2通りの手法に大別される。

ファンダメンタルズ分析

企業業績や財務諸表を元に、その企業の現在の理論株価の計算や、将来的な成長性を予測する手法

テクニカル分析

株価の一連の動きからのみ、将来の株価を予測する手法

もっともこれらの分析を行うタイミングは千差万別なので、分析結果をもとに実際に株を売買するタイミングもばらばらになる。よって機関投資家らによる大規模な取引を除いて、売買が集中しないため株価への影響も少ない。ところがひとたび新しい情報が市場に入ってきた時、ニュースの重要度が大きく、市場予想との乖離が大きくなればなるほど株の売買が集中し、株価への影響は大きい。

以上のように、ニュースが与える株価への影響は通常取引と比べ大きいと考えられる。これは、各国で禁止されているインサイダー取引に見てとれる。

2.3 日本の株式市場と指標

ある株式会社が自社の株を公開するとき、まずはじめに自国にある株式の取引所に上場することによって、全ての人がある会社の株を売買することが可能となる。日本の代表的な取引所として東京証券取引所(東証)があり、その中でも上場基準が最も厳しい市場第一部には日本の有名企業・大手企業が軒並み上場し、その数1784社となる(2014/1/9現在)。この一部に上場する銘柄の株は一般の人を含めて多種多様な人によって売買されているため、他の銘柄と比較して流動性が高く、その株価は投資家の総意を反映しやすい。

そこでこの流動性の高い東証第一部に上場している銘柄を用いた指標は以下の2つがある。

- 日経平均株価(日経平均)

日本で最も有名な指標。日本経済新聞社が定める東証一部の225社の株価平均型株価指

数である。225社の銘柄は異なる業種から銘柄をバランスよく配合して、また度々流動性などを加味して入れ替えられるため、指標として様々な人に信頼されている。日本の景況を表すときに用いられることが多い。

- 東証株価指数 (TOPIX)

東証一部に上場する全ての銘柄を用いた株価指数である。日経平均よりはマイナーではあるが、次点として用いられることが多い。東証一部全体を使うときに用いられることが多い。

図 2.2 に、過去数年の2つの指標を動きを表す。2つの指標は非常に良く似た動きをすることがわかる。

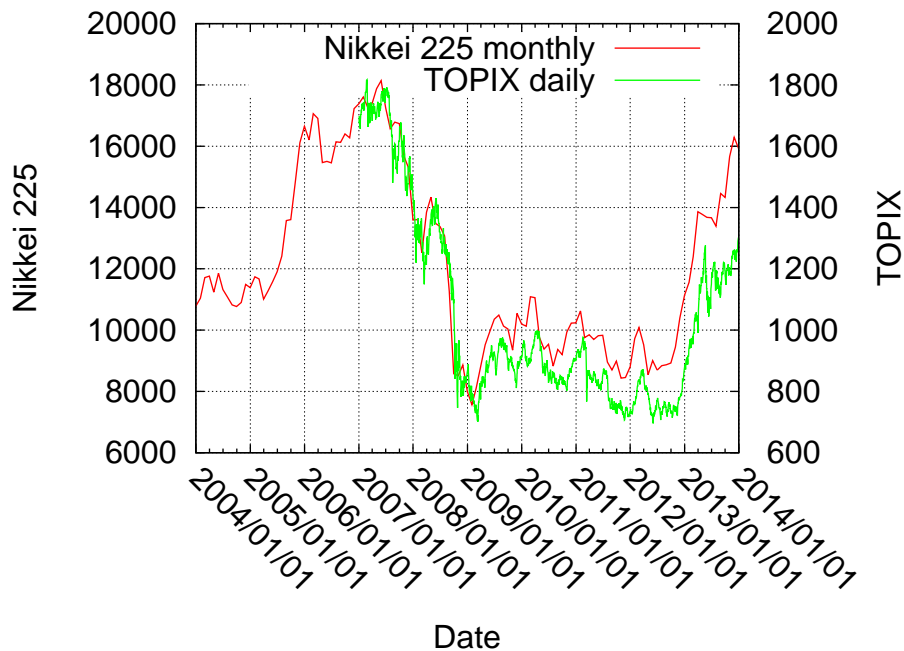


図 2.2: 日経平均と TOPIX の過去数年の推移

本研究では様々な業種からバランスよく配合されていることを考慮し、株価動向を予測する銘柄を、日経平均を構成する 225 銘柄に絞る。

2.4 外貨為替相場と株価の関係

米ドル/円などの外貨為替の相場は、輸出入を行っている企業の株価に影響を与える。例えば、ある企業が日本で製造した1万円の商品をアメリカで売るとする。もし1米ドル100円ならば、アメリカでは100米ドルでの販売となる。ここで仮に円高が進んで1米ドル80円になったとすれば、先ほどの商品は125米ドルとなり、100米ドルの時と比べ売れ行きは悪化する。そのため輸出企業の業績は悪化する。逆に円安が進み1米ドル125円になったとすれば、先ほどの商品は80米ドルとなり、100米ドルの時と比べ売れ行きは良くなる。そのため輸出企業の業績は良くなる。

輸入企業の場合は輸出企業の逆となり、円安で株価が下落し、円高で上昇する。また輸出入を行わない内需企業の株価は、一般的には外貨為替相場に影響されない。以上をまとめると、表2.1のようになる。

表 2.1: 為替と株価の関係

	円安	円高
輸入企業	下落	上昇
輸出企業	上昇	下落
内需企業	変わらず	変わらず

しかしながら現在の企業の商行為は多様化しており、必ずしも為替相場が株価に影響するわけではない。例えば企業によって輸出入を行う国が異なるためにどの通貨に対して相場が変動するかで影響のある企業が変化したり、企業によっては為替予約などの為替ヘッジを行っていることある。

実際の為替相場と株価の相関関係を見てみると、日銀が発表したレビュー [1] の中では、図 2.3 のように必ずしも為替相場と株価に相関性があるわけではないが、近年においては相関性が高まっていることが指摘されている。その相関性について詳しく分析してみると、図 2.4 のように、同時相関性と為替先行の時差相関性について正の相関(円安・株高、円高・株安)が見られることも指摘されている。その背景として、上記で説明したことはもちろんのこと、海外投資家の売買比率が高まっていること、為替と株価の相関関係を利用した高速・高頻度のプログラム売買 (High Frequency Trading, HFT) が少なからず存在している可能性について言及している。レビューの最後には、過度な相関性によるボラリティの発散によるリスクの増大の危険性についても触れている。

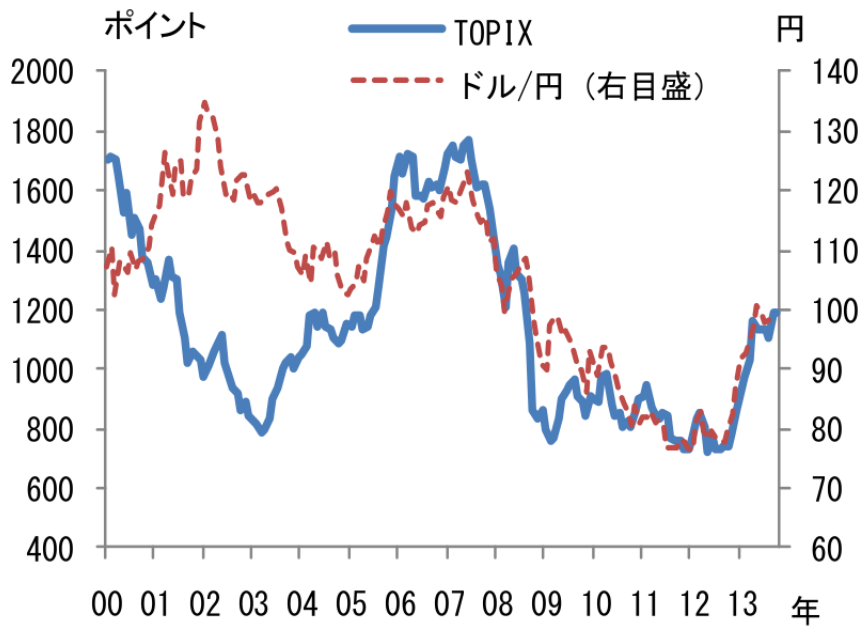


図 2.3: TOPIX と米ドル/円の推移 [1]

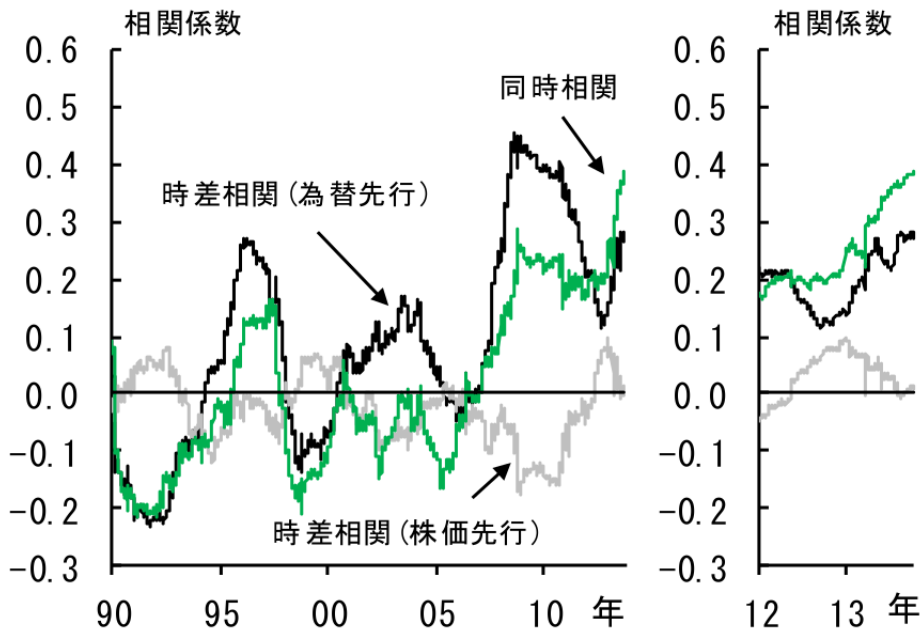


図 2.4: TOPIX と米ドル/円の相関性 [1]

第3章

テキストマイニング

3.1 データマイニング

テキストマイニングの上位概念であるデータマイニングは、特定の問題を対象とせず、特に大量のデータから統計学や機械学習などの様々な技術を用いることによって、通常の分析からは得られないような知識を獲得することを目指す。このような技術が発達した背景には、近年の高度情報社会化が大きく寄与しており、インターネットの通信速度や各端末のストレージ容量が飛躍的に増大したため、各人が扱えるデータ量が爆発的に増え、データマイニングによって新規の知識を獲得することが容易になったためである。そこでインターネット上に溢れる大量のデータを用いてデータマイニングする試みが一気にメジャーになった。

そしてテキストマイニングは、テキストデータを対象としたデータマイニングを表す。データマイニングの下位概念としてテキストマイニングという概念がある背景には、テキストデータは他の音声データなどに比べて機械的に扱いやすく、ウェブ上にテキストデータが大量にあり、それ自体は他と比較してデータ容量が小さいため、他のデータを対象としたデータマイニングと比べて普及しているためである。

データマイニングは、大きく以下の4つのステップに分かれる。

1. データ収集

ウェブ上ないしはその他メディアより、テキストデータや画像データ、音声データ、映像データなど目的に沿ったデータを収集する。

2. データベース化

集めたデータをデータベースに投入する。

3. 解析

統計学や機械学習を用いて、集めたデータに関する解析を行う。

4. 知識獲得

解析結果からデータに関する知識を獲得するが、必ずしも目的が達成されるわけではない。

ここで目的の成否を決める重要となるのは、最初と3番目のステップである。最初のデータ収集のステップで集めるデータに獲得したい知識が含まれていなければ、いくらデータマイニングしても目的を達することはできない。ただし集めたデータに目的の知識があるかどうかは解析するまではわからないことが多い。3番目の解析のステップでも、その手法次第では集めたデータから限らない知識を得ることも可能であるが、逆に手法がマッチしなければ知識を獲得することは困難である。総じて言えば、データを活かすも殺すも解析手法次第

である．そこでデータマイニングで重要となる，解析手法について次節で説明する．

3.2 データマイニングにおける解析手法

まず，集めたデータからどのような問題(タスク)を解くかによって，以下のように分かれる．

- クラスタリング
似ているデータ同士を同じグループに集める．予めグループ分けに関する情報与えるのは，必要最小限に留める．
- クラス分類
データが振り分けられるべきクラスの全てを予め定め，データをそれぞれのクラスに機械的に分類していく．
- 系列ラベリング
品詞のタグ付けに代表されるように，ある系列に対して順次クラス分類を行う．クラス分類と似ているが，同じ系列に対して以前の分類結果を利用する点で異なる．
- 回帰分析
データと目的の変数の間の関係を学習し，目的の変数を予測する．
- パターン認識
画像データに含まれる物体を認識するなど，データから一体のパターンを抽出する．

いずれもデータマイニングによって良く使われるが，この中でもテキストマイニングで良く使われるのがクラス分類である．例えば与えたメールが迷惑メールかどうかを判定する迷惑メールフィルタは，代表的なクラス分類問題である．

本研究でもクラス分類問題をメインのタスクとして扱うため，クラス分類問題でよく使われる機械学習の手法について説明する．大きく以下の手法が存在する．

- ナイーブベイズ分類器
統計学的にデータを分類する手法．古典的ではあるが，計算量が少なく現在でも使われることがある．
- サポートベクターマシン (SVM)
高次元空間において，データを2クラスに分類する分離平面を求める手法．データ量に対して計算量が指数的に増加する次元の呪いにも強い．汎用性の高さで性能の高さから

最も良く使われる手法であるといっても過言では無い。

- ニューラルネットワーク

人間の脳内ネットワークを模した手法。人間の学習アルゴリズムに近いので、使い方が次第であらゆる学習をさせることができ、その汎用性の高さはSVM以上であるが、必ずしも高い性能を発揮するとは限らない。SVMの登場によってその人気も下火であったが、近年のDeep Learningアルゴリズムの登場によって再度人気を博している。

どの手法を採用するかは好みのものであるが、現在ではSVMとニューラルネットワークが主流の手法である。そこで本研究では、学習効率の良さも加味してSVMを選択した。

3.3 サポートベクターマシン

サポートベクターマシン(SVM)[4]は前述の通り、与えられた特徴空間に写像されたデータを2クラスに分類する分類器である。その手法自体はとてもシンプルで、空間に存在するデータを最も正しく2つの空間に分離するような超平面(分離平面)を求めるだけである。2次元空間の場合は図3.1のようになるが、多くの場合はもっと高次元である。ただ闇雲に分離平面を引いても分類性能は高くないが、その手法としての最大の特徴は、マージン最大化とカーネル法にある。

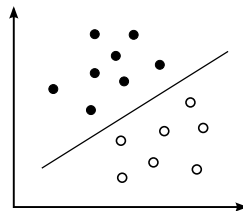


図 3.1: SVM による 2 次元空間の分離平面の例

マージン最大化

マージン最大化とは、図 3.2 の A の分離平面のように、分離平面から最も近い2つのクラスに対して、その距離(マージン)が最も大きくなる分離平面を求めることを指す。B の分離平面は、分離平面から近いデータが A のそれよりも近い。C の分離平面は、マージンこそ大きいですが、そもそも正しく分離できていない。このように分離平面のマージンを最大化することによって、未知のデータに対しても正しく分離できる可能性が高いのである。こ

のことを汎化能力が高いという。

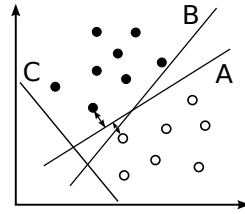


図 3.2: SVM による様々な分離平面

カーネル法

カーネル法は、もともと線形にしか分離できない SVM に、カーネル関数によって特徴ベクトルをより高次元の空間に写像することで線形に分離できるようにし、もとの空間における非線形な分離を可能とする方法である。例えば図 3.3 のような線形分離できないデータに対して、カーネル法によって図 3.4 ような空間に写像することによって、線形分離が可能となる。詳しい説明は数学的な話が多くなってしまいうため割愛するが、カーネル関数はどのような関数でもなれるわけではなく、計算量を抑えるために一定の条件を満たす必要がある。逆に言えば、カーネル法を使っても線形 SVM と比較して計算量が指数的に増えるわけではないので、SVM としては通常カーネル法を用いた非線形 SVM が用いられる。

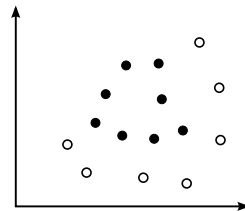


図 3.3: SVM によって線形分離できない例

3.4 テキストマイニング特有の解析手法

データマイニングで頻繁に使われる機械学習、特に SVM について説明したが、この節ではテキストマイニングに特有な解析手法について説明する。

まず、前述のように SVM では特徴空間に写像されたデータを分類する分類器であるが、このデータのことを特徴ベクトルと呼び、1 つ 1 つのスカラ値を特徴量と呼ぶ。すなわち、

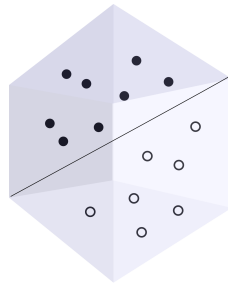


図 3.4: カーネル関数によって線形分離できる空間へ写像

集めたテキストデータを何らかの手法によって特徴ベクトルに変換しなければいけない。テキストデータから特徴ベクトルに変換する代表的な手法として、Bag-of-Words がある。

Bag-of-Words (BoW)[5] とは、テキストデータの中で出現する単語の数を数え、各々の単語の出現回数をベクトルとする手法である。ここでベクトル上のあるスカラー値は、どのテキストデータに対しても同じ単語の出現回数に対応する必要があり、また他のテキストデータに出現している単語でもあるテキストデータに出現しなかった場合は、その単語の値を省略せず 0 としなければいけない。

例えば、

This is a pen.

と

The man with a pen.

という文章をそれぞれ BoW で特徴ベクトルに変換する場合、前者は

$$\begin{bmatrix} \text{this} & 1 \\ \text{is} & 1 \\ \text{a} & 1 \\ \text{pen} & 1 \\ \text{the} & 0 \\ \text{man} & 0 \\ \text{with} & 0 \end{bmatrix}$$

となり、後者は

this	0
is	0
a	1
pen	1
the	1
man	1
with	1

となる。

このような短文を BoW に変換した場合でも特徴ベクトルの次元は 10 程度になるが、たくさん長い文章を BoW に変換した場合その次元は万のオーダーを超えることも珍しくない。そのためテキストマイニングでは次元数が極端に大きくなるので、次元の呪いの影響を受け、指数的に計算量が爆発する。

そこで次元数を削減するために、ストップワードと呼ばれる、自然言語処理において排除しても影響のない一般的な単語を除去する必要がある。先ほどの例で言えば、“the”, “is”, “a”, “this” が当たる。ストップワードを除去することによって、性能を落とさずに計算量を減らすことができるようになる。ただしストップワード自体は固定的であり、全体の語彙数からその割合は僅かなものなので、ストップワードを削除しても格段に性能が向上し、計算量が減るものではない。

3.5 日本語固有の問題

テキストマイニングする場合は、対象となる言語によって手法が一部変化する場合がある。本研究では日本語を対象としたテキストマイニングを行うが、日本語固有の事情として、英語などと異なり単語間に明確な区切りが無いことが挙げられる。このため英語のようにスペースで区切って BoW にするという手法が使えず、日本語の場合何らかの手段を用いて文章を分かち書きし、単語に分解する。例えば、

すもももももものうち

という難解な日本語は、どこが単語の区切りかは機械的に判断することはできず、構文解析を行う必要がある。更に分かち書きを行うのと同時に、その単語の品詞を推定する必要もあり、このような分かち書きと品詞推定を行うプログラムを形態素解析器と呼ぶ。

本研究ではこの形態素解析器として、実績もあり性能も安定している MeCab[6, 7] を利用

した．また合わせて解析に必要な辞書としては，推奨されている IPA 辞書を利用した．この辞書を変えると解析結果が変わるのはもちろん，品詞の分類方法も変わるので注意する必要がある．

ちなみに先ほどの例文を MeCab を用いて解析を行うと，

すもも	名詞, 一般, *, *, *, すもも, スモモ, スモモ
も	助詞, 係助詞, *, *, *, も, モ, モ
もも	名詞, 一般, *, *, *, もも, モモ, モモ
も	助詞, 係助詞, *, *, *, も, モ, モ
もも	名詞, 一般, *, *, *, もも, モモ, モモ
の	助詞, 連体化, *, *, *, の, ノ, ノ
うち	名詞, 非自立, 副詞可能, *, *, *, うち, ウチ, ウチ

という結果が得られる．これを BoW にすると，

すもも	1
も	2
もも	2
の	1
うち	1

という結果が得られる．

ただし日本語の場合，同じ単語の形でも文脈によっては品詞が異なり，意味が違う可能性があるため，BoW にする場合は同じ単語の形かつ同じ品詞のものを同一の単語として数える必要がある．例えば「あげる」という単語には非常に多様な意味が存在し，

行ってあげる

の場合，品詞は動詞の非自立となるが，

高くあげる

の場合，品詞は動詞の自立となり，同じ単語の形でも意味が異なる．このように単純に同じ単語の形のものと同じ単語として数えると，意味が異なる単語を同じものとして扱ってしまう可能性がある点に注意する必要がある．

第4章

提案手法

この章では、既存研究における問題点とそれに対する本研究のアプローチを述べ、その後提案手法の全体像を俯瞰したあと、詳細について説明する。

4.1 既存研究の問題点

既存研究では、主に以下の問題点が指摘される。

- 新聞媒体のニュースに対するテキストマイニング
- 制限のないニュースの対象化
- 確立したテキストマイニング手法が存在しない

以降の各節で、問題の詳細とそれに対する本研究のアプローチについて述べる。

4.1.1 新聞媒体の問題

新聞媒体のニュース記事を対象とし、それらをテキストマイニングし、株価などの何らかの金融指標を予測する研究 [8] が過去に行われてきた。

一般的に投資家らは証券会社などが提供する専門のニュース媒体を利用することで、鮮度の高いニュースを仕入れている。そのためイベントが発生した場合に、投資家らはすぐにその情報をキャッチし、取引が行われることで株価は即座に反応する。

しかしながら、一般的な朝刊としての新聞は、最悪1日遅れで情報が伝わる可能性があり、鮮度に欠けることになる。このようなニュースをテキストマイニングしても、その情報の価値は既に株価に織り込まれているため、予測する余地が存在しないことになる。

より鮮度の高いニュースとして、ウェブ上のニュースを扱った研究 [9] もあるが、それでも投資家らが使うようなニュース媒体ではなく、あくまで全国紙などの庶民的な新聞である。このような新聞ソースのニュースは、速報性・専門性に欠け、また発生した金融イベントの大部分が記事にならない可能性が高い。そのため新聞ベースのソースを金融テキストマイニングで扱うには適さない。

そこで本研究では、投資家らを対象としたウェブ上のニュースソースをテキストマイニングの対象とすることで、ニュースの鮮度を確保する。具体的には、証券会社が配信している顧客向けのニュースを利用する。¹

¹証券会社がニュースを提供しているわけではなく、あくまで専門の会社から提供されたニュースを配信しているだけである

4.1.2 制限のないニュースの対象化

既存の研究では、あるニュースソースからニュースを取得したら、取得できたニュースの全て、もしくは会社名などによるフィルタリングされたニュースを用いてテキストマイニングしていた。しかしながら、全てのニュースに必ずしも情報としての価値あるとは限らないことが問題として指摘できる。

まず、一般的にはニュースの種類として以下の3つに分類できる。

- 過去に発生したイベントのニュース(うち、速報と非速報の2つに分けられる)
- 何らかの現在の状態をまとめたニュース
- ある事柄に関する未来を予測するニュース

このうち、株価に直接的に影響を与えやすいのはイベントの速報ニュースであるが、このようなニュースは数としては多くない上に不定期であるため、テキストマイニングの対象としては適さない。

次に情報としての価値が高いのは、その情報が織り込まれない可能性が高い、未来を予測したニュースであるが、反面その確度に関しては著者の予測能力に依存するため、ニュースから得られる情報が実際に起こるとは限らないため適さない。

最後に現在の状態のまとめたニュースであるが、一般的にはこのようなニュースは市場で既に織り込まれていることが多いが、場合によっては長いスパンで見た現在の状態に関するニュースもあるため、そのようなニュースはこれから市場で徐々に織り込まれる。

そこで本研究ではこのようなニュースとして、為替相場に関する市況を扱ったニュースをテキストマイニングの対象として絞った。為替市況のニュースは、直近の為替相場の動向とそれに関する原因について触れるため、そこから導き出される直近の未来を予測することが可能であると考えたためである。

4.1.3 確立したテキストマイニング手法の非存在

従来の研究では、テキストマイニングの手法として、特徴量の設計や機械学習の手法の選択など、様々な面で研究者依存な部分が多かった。これは確立したテキストマイニング手法が存在しないことに起因する。また一部言語依存な面もあるため、他言語の研究をそのまま適用できるとは限らない。

そこで本研究では、既存研究の手法を参考にしつつも、できる限り言語に依存しない手法を提案することで、この問題に対処する。

4.2 提案手法の全体像とアルゴリズム

提案手法の全体の流れは図 4.1 のようになる。

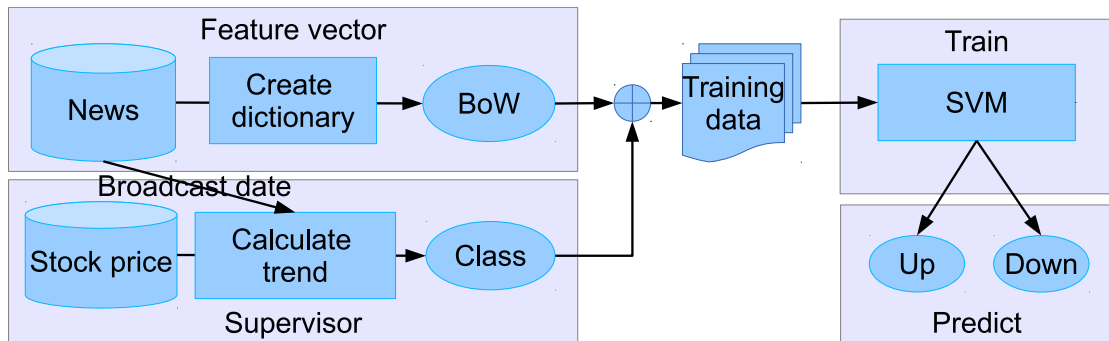


図 4.1: 提案手法の全体図

アルゴリズムとしては、以下の通りになる。

1. 全記事に対して、BoW の辞書を作成するために以下を繰り返す
 - 1.1. MeCab で分かち書き
 - 1.2. 品詞フィルタリング
 - 1.3. 辞書に出現回数を登録
2. 辞書の中から、下限より低いまたは上限より高い出現頻度の単語を除去する
3. 銘柄毎に、学習データを作成するために全記事に対して以下を繰り返す
 - 3.1. 教師信号として、記事の発表時刻から見て x 分後の株価が上昇・下降しているか調べ、もし変化していなかったらこの記事の学習データは作成しない
 - 3.2. 記事を MeCab で分かち書きをした後、辞書に無い単語を除去する
 - 3.3. 残った単語と出現回数を、辞書の出現順に並べ BoW を作り、学習データに加える
4. 銘柄毎に SVM に学習させ、モデルを作り、評価する
 - 4.1. 訓練セットとテストセットをスケーリングする
 - 4.2. 訓練セットに対して、5 分割交差検定で最も成績の良い SVM のパラメータを探索する
 - 4.3. 訓練セットに対して、上記のパラメータで SVM の学習モデルを作成する
 - 4.4. テストセットに対して、学習モデルで予測した結果と比較して評価する

以降の節で、各ステップの詳細を説明する。

4.3 自然言語処理

この節では、本研究で適用した自然言語処理の手法について説明する。

4.3.1 単語のフィルタリング

SVM に投入する記事の特徴量として、各記事の BoW を作成する必要があるが、単語によってはフィルタリング対象として除去される可能性があるため、まず最初に BoW の元となる辞書を作成する。

そのため一旦全テキストを MeCab にて形態素解析を行ったあと、単語単位のフィルタリングを行いつつ、辞書を作成する。その単語のフィルタリング条件としては、以下の2つがある。

- 品詞
- 単語出現頻度

これらのフィルタについて、詳細は以下で説明する。

品詞フィルタ

まず品詞によるフィルタリングに関しては、Schumaker らの研究 [10, 11, 12] でもその有効性が確認されているが、対象とする言語が違うために、本研究ではフィルタリング条件を緩くすることにした。条件としては、「単体で意味を成さない単語」とする。例えば「に」や「は」などの助詞がそれに当たる。これらの単語を除去することによって、学習の妨げとなるノイズを特徴量から減らすことができる。

MeCab+IPA 辞書の場合、付録 A のように細かく品詞が付与されるが、本研究では表 4.1 の品詞を除去した。

単語出現頻度フィルタ

最後に、辞書が出来上がった後に単語の出現頻度のフィルタリングとして、出現頻度に閾値として下限値と上限値を設定し、下限値未満もしくは上限値を超える単語を除去する。単語 w の出現頻度 f_w は、 N を記事数、 w_i を i 番目の記事における単語の出現回数とし、式 4.1 のように計算する。

表 4.1: 本研究でフィルタリングする品詞

品詞	細分類 1	細分類 2	細分類 3	例
感動詞	*	*	*	すいません
記号	*	*	*	()「」,..
助詞	*	*	*	に,は
助動詞	*	*	*	です,ます
接続詞	*	*	*	だけど
名詞	数	*	*	0, 1, 99
名詞	副詞可能	*	*	全て
名詞	接尾	助数詞	*	円
名詞	非自立	*	*	はず

$$f_i = \frac{\sum_i^N w_i}{N} \quad (4.1)$$

出現頻度が低い単語はその単語の存在がクラス分類に影響を与える可能性は低く、逆に出現頻度が高い単語はクラスに関わらず出現すると考えられ、どちらも予測に適さない。よってこれらの単語を予め除去することによって、BoWのノイズを減らすことが出来ると考えられる。

この手法は、高出現頻度の単語はストップワード除去の同じ効果があると考えられるので、従来言語別にストップワードを定義して除去していたのが必要なくなる。事実、本研究ではある程度は品詞フィルタリングは行っているものの、単語単位でのフィルタリングは行っていないため、言語非依存な手法であると言える。

4.3.2 入力テキスト

ここまで自然言語処理としての単語のフィルタリングについて述べたが、そもそもシステムに入力するテキストとしてニュースの記事本文を用いると、そのニュースの本質となる重要な文章が、他の重要でない文章と同等の扱いを受けるため、BoWによって特徴ベクトルに変換したときに埋もれてしまう可能性がある。

そこでニュースの記事本文ではなく、タイトルを用いることによってこの問題に対処する。何故ならばニュースのメインピックスは、必ずタイトルに含まれるからである。ただし記事本文に含まれる、メインピックスをサポートするサブピックスも捨てられてしまうの

で、この点は問題として残ることになる。この問題については、第7章の今後の展望でも触れる。

4.4 教師信号

本研究ではニュース記事を元に株価動向を予測するので、株価動向が教師信号となる。この節では、その教師信号に関する手法について述べる。

4.4.1 分類クラス

株価動向を自然に考えるならば、分類されるクラスとしては上昇・変化なし・下降の3クラスになる。ところが一般的に株価は常に動いていて、全く変化していないということは稀なので、上昇・下降の学習データに対して、変化なしの学習データは極端に少なくなる。このような不均衡データの場合、SVMでは少ない方の学習データが学習しにくく、多い方の学習データを予測する傾向にある。

そこで本研究では、教師信号が変化しなかった場合の学習データは除去し、上昇・下降の2クラス分類問題に落としこむことにする。こうすることのメリット・デメリットは以下のようになる。

- メリット
 - 均衡データに近づくので、SVMの学習性能が向上する
 - 2クラスなので結果を評価しやすくなる
- デメリット
 - 学習データが少なくなる
 - 実際に予測する時に、変化しないという予測ができなくなる

株価が変化しないことは稀であるということを考慮すれば、デメリット自体は極僅かであると考えられる。

4.4.2 株価動向の定義

前項でも、株価は常に動いているということを述べたが、その動きは時に激しく動くこともあり、ある時点の株価はその前後の株価と比較して大きく異なる値となっていることもしばしばある。そのため、ある時点の株価と x 分後の株価を比較して単純に大きくなっていれば上昇、小さくなっていれば下降とすることは実際の株価動向にそぐわない可能性がある。

そこである時点 t の株価を P_t とし, x 分後の株価変化率 $rate$ を式 4.2 で求める.

$$rate = \frac{\sum_{i=0}^4 P_{t+x+i}/5 - \sum_{i=0}^4 P_{t-i}/5}{\sum_{i=0}^4 P_{t-i}/5} \quad (4.2)$$

この式はつまり, ある時点 t の株価 P_t を, t の時点を含む過去 5 分間の平均株価とみなし, 時点 t から x 分後の株価 P_{t+x} を, $t+x$ の時点を含む向こう 5 分間の平均株価とみなした時の, 株価変化率である. これを参考までに図示したものが, 図 4.2 となる. この図のように, 全体のトレンドが下降傾向なのに, たまたま大きく上昇したために, x 分後の株価が上昇していると誤認してしまうことを減らすことができる.

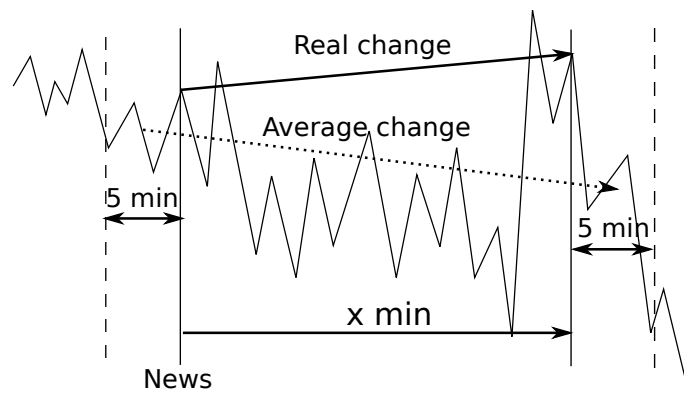


図 4.2: 株価動向の求め方のサンプル

ここでなぜ平均株価を求める際に, ある時点の前後 5 分間の平均株価ではなく, ある時点から過去または未来の 5 分間の平均株価としたかを説明する. もしニュースが配信されたからの前後 5 分間の平均株価を採用した場合, 僅かながらとはいえ, ニュースが配信された影響による株価変動を取り込んでしまう可能性があるためである. x 分後の前後 5 分間ではなく, 向こう 5 分間の平均株価としたのも, x 分の値が小さい時に, ニュースの影響で株価が安定していない可能性があることを考慮したためである. とはいえ, 結果としては前後 5 分間の株価でもあまり変わらない可能性はあるが, 本研究ではこの方法を採用した.

4.4.3 気配値

前項では株価変化の定義をしたが、この項ではそもそも株価として値が見つからない場合を紹介する。

ある企業の株価というのは、その株に対して最も最近取引された金額を言う。つまり取引が無ければ、株価は更新されず古い状態が維持されるが、市場が開いた瞬間これはリセットされる。つまり市場が開いて初めて売買が成立した瞬間に、その日の最初の株価は決まるのである。

ところが例えば、ある企業の株に対して市場が開く前から売り注文が殺到し、買い注文が無かったとすると、市場が開いても売買は成立しない。このような場合、売買は成立していないので株価としては値が付かない状態である。

このように株価が表示されていない状態でも、その投資家へ目安として気配値が提示されるが、この情報というものは一般的に保存されないのので、最終的には株価として値が付かなかったという情報のみが残る。

そこで本研究ではこのような場合に対して、ある時刻 t の株価 P_t を求める際に、その時点で値がついていなかった場合、 t を過去に遡り、最も近い時点で値がついていた時の株価を時刻 t の株価とする。

例えば前日の終値が 1000 円の株があったとき、9 時に市場が開いてから 5 分間値が付かず、5 分後に 1100 円の値が付いたとする。このような場合、9 時 5 分までの株価を 1000 円とみなし、9 時 5 分以降を 1100 円とみなす。

4.5 SVM による学習

この節では、本研究で用いた SVM による学習と評価の手法について説明する。

4.5.1 学習データのスケールリング

SVM による学習を行う前に、学習データのスケールリングを行う必要がある。これは仮に出現回数が非常に多い単語があった場合に、その単語が他の単語と比べて重みが大きくなってしまうためである。

本研究での特徴量は単語の出現回数なので、 $[0, \infty]$ を $[0, 1]$ にスケールリングする。

4.5.2 SVM ライブラリとパラメータ

SVM のライブラリとして出回っているものは様々あり、用いるライブラリによってパラメータや細かい部分の手法に差異があるため、分類結果が異なることが多い。そこで本研究では libsvm[13, 14] を使用することとした。

libsvm のパラメータとしては、結果に大きく影響するものとして以下の3つがある。

- カーネル
線形, 多項式, RBF(デフォルト), シグモイドの4種類
- コストパラメータ
分離の失敗に対する厳しさを表すパラメータ。これが小さいほど失敗を許容し、大きいほど失敗を許さない。
- カーネル関数のパラメータ (線形カーネル以外)
多項式は3つ, RBFは1つ, シグモイドは2つ

libsvm の作者である Chang 氏によると、カーネルは大抵の場合 RBF が最も優れているためデフォルトのままで大丈夫であるが、コストパラメータとカーネル関数のパラメータは最適なものを探索する必要がある [15]。そこで本研究でもカーネルは RBF を用いることとし、その RBF 関数のパラメータ γ とコストパラメータの2つについて、最適なものを探索することとする。

探索方法としては、作者が推奨しているグリッド探索を用いることとする。グリッド探索の範囲についても、デフォルトで指定されている、コストパラメータは $2^{-5+2x}; 0 \leq x \leq 10$ の範囲で、 γ パラメータは $2^{-15+2x}; 0 \leq x \leq 9$ の範囲で探索する。

また libsvm には、各クラスのコスト (ペナルティ) の重み付けを調整することで、ある程度の不均衡データに対応することができる。² 例えば、正例のデータ数が少ない場合は正例のコストを大きくすることで、負例寄りの予測をする不適切な学習を減らすことができる。そこで本研究でも、この調整を行うことにする。具体的には、正例と負例のデータ数をそれぞれ $p, n(p < n)$ とした場合、それぞれのコストは $C_p = p/p = 1, C_n = p/n < 1$ で重み付けする。

²ただし限界はあるので、これに頼らずデータ数を調整することが望ましい。

4.5.3 n 分割交差検定

前項のパラメータ探索時に、学習データに対して過度に適応してしまい、未知のデータに対する識別能力が下がってしまう過学習という現象が、どのような機械学習でも発生する可能性がある。そこで通常は訓練用の学習データ(訓練セット)とは別に検証用の学習データ(検証セット)を用意し、検証セットに対して評価することで過学習を防ぐ。ところが検証セットにデータを割く分、訓練セットの学習データの数が減ってしまうデメリットがある。そこで検証セットを別に用意することなく、訓練セットだけで過学習を防ぐための手法として、 n 分割交差検定を本研究では用いた。

n 分割交差検定は、図 4.3 のように学習データを n 個のブロックに分割し、ある 1 つのブロックを検証セットとみなし、残りのブロックで学習モデルを訓練した後、検証セットで評価するのを 1 つのステップとする。このステップを順次全てのブロックについて繰り返したあと、最終的に全ての評価結果を平均する手法である。この手法を用いることによって、別途検証セットを用いず訓練セットの学習データだけで過学習を防ぐことができる。 n の値は研究によって様々であり、またその値についても根拠がないことが多いが、本研究ではデータが少ない場合分割されすぎることによる学習データ不足を防ぐために、5 分割交差検定を採用した。

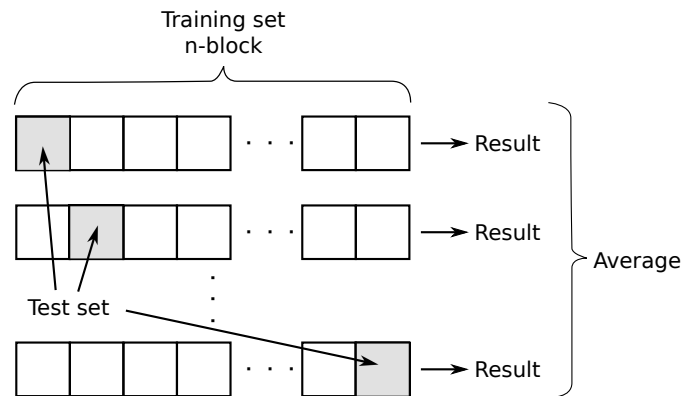


図 4.3: n 分割交差検定

4.6 評価方法

学習したモデルによるテストセットでの予測結果に対して、評価方法としては以下の 4 つを用いる。

- 訓練セット正答率
- テストセット正答率
- 平均 F 値
- 仮想トレードのリターン

4.6.1 正答率

訓練セット正答率およびテストセット正答率は、単純に訓練セットおよびテストセットのそれぞれについて学習したモデルによる予測精度を表す。厳密には訓練セット正答率はパラメータ探索に用いるためのもので評価用ではないが、参考データとして付記する。

4.6.2 平均 F 値

既存研究では主に予測の正答率(精度)によって、結果を評価することがほとんどであった。しかしながら精度だけでは評価することができない部分がある。それが予測の偏りである。

例えば9割が正例のデータがあったとする。このとき常に正例を予測すれば、実に9割の精度を得ることができる。ところが負例のみの精度は0である。このように一見精度が高いように見えても予測に偏りが見られる場合があり、精度からでは予測の偏りを評価することができない。

そこで、本研究では予測の偏りの尺度として、文書分類の研究で使われる F 値を導入する。しかしながら文書分類と本研究とではタスクが異なり、単純に F 値を用いると評価尺度としての都合が悪いので、平均 F 値を求めることにする。

平均 F 値は各クラスの F 値のマクロ平均、すなわち正例と負例のそれぞれで F 値を計算し、その平均をとることにする。これは正例の F 値のみの場合、正例と負例が同確率だと仮定するならば予測を全て正例とする分類器でも高い F 値となるためである。

F 値の計算式については、付録 B に記載するが、例えばとあるテストセットで表 4.2 の結果が得られたとする。

表 4.2: 2 クラス分類で得られた結果の例

予測 \ 実際	正例	負例
正例	40	40
負例	10	10

この時、正答率は 50% となるが、平均 F 値は以下のように計算される。

$$\begin{aligned} \text{正例の適合率} &= \frac{40}{40+40} = 0.5 \\ \text{正例の再現率} &= \frac{40}{40+10} = 0.8 \\ \text{正例のF値} &= \frac{2 \cdot 0.5 \cdot 0.8}{0.5+0.8} = 0.6154 \\ \text{負例の適合率} &= \frac{10}{10+10} = 0.5 \\ \text{負例の再現率} &= \frac{10}{10+40} = 0.2 \\ \text{負例のF値} &= \frac{2 \cdot 0.5 \cdot 0.2}{0.5+0.2} = 0.2857 \\ \text{平均F値} &= \frac{0.6154+0.2857}{2} = 0.4506 \end{aligned}$$

このように、偏った予測をすると平均F値は悪化するので、平均F値を見ることで偏った予測をしているのかがわかる。

株価が上昇する確率と下降する確率がそれぞれ等しいとし、それをランダムに予測したのならば、テストセット正答率と平均F値はそれぞれ0.5となるので、これがベースラインとなる。

4.6.3 仮想トレード

最後に仮想トレードのリターン(収益率)は、提案手法にて取引を行ったと仮定した場合の収益率である。本研究では現在の資金に関係なく、常に一定金額分の売買を行い、取引に関わる全ての手数料は考慮しない。つまり100万の資金からスタートし、1万円の株を100万円分買い、その後株価が1万100円になった時に全株を売れば101万円になるが、次の取引でも売買するのは100万円分なので、全取引のリターンは個別の取引のリターンの総和となる。そして取引のタイミングは、ニュース配信時に売買を行い、それから予測期間後に反対売買が成立したと仮定した。なおこの評価方法は最終的に提案手法が実取引に通用する可能性があるかどうかを見極めるために用いるので、全ての実験ではなく、最終評価に用いる。

第5章

株価動向予測実験

実験では、特定の銘柄の株価を様々な予測期間で予測し、その結果を評価する。まずベースラインとしてランダム予測した場合と BoW のみで予測した場合の結果を評価し、その後各種提案手法を適用した場合の結果を評価する。最後に、最も評価が良かった結果の詳細を述べ、分析する。

5.1 データ

5.1.1 ニュースデータ

ニュースデータとして、日経 QUICK ニュース社より平日の 8 時半、10 時、12 時、17 時に配信されている為替市況のニュース¹を用いた。ただし 17 時のニュースは東証の立会時間外なので用いない。²

5.1.2 株価データ

株価データは、日経平均を構成する 225 銘柄の中から輸出企業・輸入企業・内需企業のそれぞれを 5 社ずつ、表 5.1 の 15 社の銘柄を適当に抽出し、1 分足のものを取得して用いた。輸出入・内需の別は、一般的に言われている業種による分類であり、個別の企業が実際に輸出入をしているかどうかは考慮しない。なお、東証の開いている時間は 9 時から 15 時（うち 11 時半から 12 時半までは昼休み）のため、必然とこの時間の間の 1 分足のデータになる。以降の実験では、同じパラメータに対して 15 社それぞれで実験を行い、結果を平均したものをそのパラメータにおける結果とした。

5.1.3 データ期間

上述の為替市況のニュースが配信開始されたのが比較的新しいため、データ量としては十分ではないが、訓練セットとして 2012 年 11 月から 2013 年 11 月まで、テストセットとして 2013 年 12 月のデータを用いた。

そのため実際の学習データ数は訓練セットとテストセットのそれぞれが 631 個と 60 個となるが、これはあくまで最大数であり、同じニュースでも銘柄によっては株価変化が無かったとして学習データから除去される可能性があるため、この数よりは幾らかは減ることとなる。

¹必ずしもその時間に配信されるわけではなく、通常 10～20 分程度遅れる

²厳密には 8 時半も時間外であるが、30 分後の 9 時より開くので採用した

表 5.1: 実験に用いた銘柄一覧

銘柄	証券コード	業種	種別
トヨタ自動車	7203	輸送用機器	輸出企業
ソニー	6758	電気機器	
ニコン	7731	精密機器	
ファナック	6954	電気機器	
ブリヂストン	5108	ゴム製品	
丸紅	8002	卸売業	輸入企業
東京ガス	9531	電気・ガス業	
東京電力	9501	電気・ガス業	
JX 日鉱日石エネルギー	5020	石油・石炭製品	
日清製粉グループ	2002	食料品	
鹿島建設	1812	建設業	内需企業
三菱 UFJ ファイナンシャル・グループ	8306	銀行業	
東日本旅客鉄道	9020	陸運業	
セブン & アイ・ホールディングス	3382	小売業	
東京ドーム	9681	サービス業	

参考までに、2013年の日経平均の推移は図 5.1 となる。アベノミクス効果で前半までは上昇相場であるが、その後乱高下を繰り返すボックス相場となっている。

5.2 実験方法

提案手法においては、予測期間と単語出現頻度の下限値・上限値の3つの値がパラメータとして残っている。これら3つのパラメータを同時に探索する場合、結果を視覚的に表しにくいのと、予測期間と単語出現頻度で従属関係が無いいため、それぞれを独立したパラメータとして考える。そのため本実験では、以下の2種類の実験を行う。

- 予測期間をパラメータとする実験
- 単語出現頻度の下限値と上限値をパラメータとした探索(予測期間固定)

基本的には予測期間をパラメータとし、単語出現頻度の下限値・上限値は固定とするが、単語出現頻度によるフィルタリング効果を検証する実験では、予測期間を固定し単語出現頻度を変化させる。

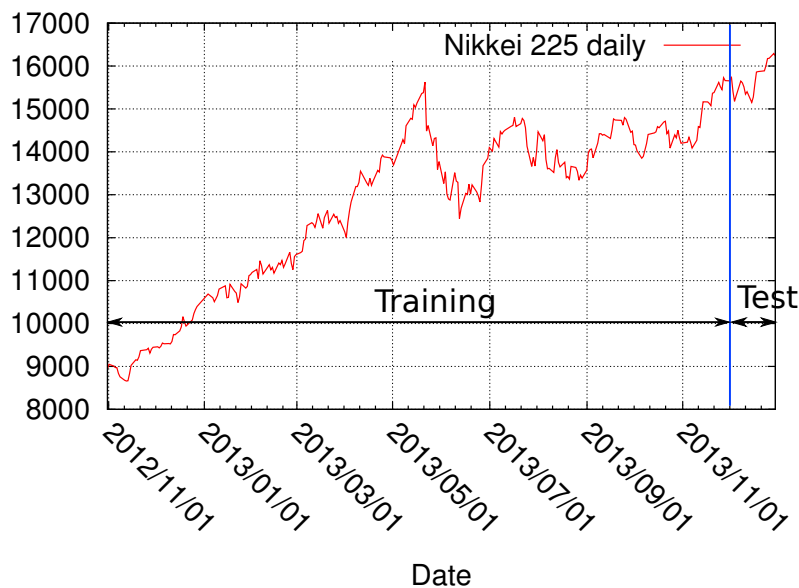


図 5.1: 2013 年の日経平均の推移

5.3 ベースライン

まずベースラインとして、ランダム予測と BoW のみで予測した場合を評価する。

5.3.1 ランダム予測

ランダム予測で予測期間を変化させた実験の結果は、図 5.2 のようになった。多少の上下はあるものの、テストセット正答率や F 値は平均してほぼ 0.5 であった。これは理想的なデータに対するランダム予測した場合の理論値と近く、このことから学習データに偏りが少ないことが言える。

5.3.2 Bag-of-Words のみで予測

BoW で特徴ベクトルに変換した後フィルタリングなどを一切行わずに、予測期間を変化させた場合の実験結果は、図 5.3 のようになった。

結果としては、テストセット正答率はほぼ 0.52 から 0.56 の間に収まって、ランダム予測した場合と比較して大きく上昇している。逆に F 値はほぼ 0.46 から 0.48 の間で、大きくても 0.5 程度と、ランダム予測の場合より減少している。正答率が改善したのは、用いたデータが為替市況のニュースであることが要因の 1 つと考えられるが、F 値が低いため予測に偏りがあると言える。

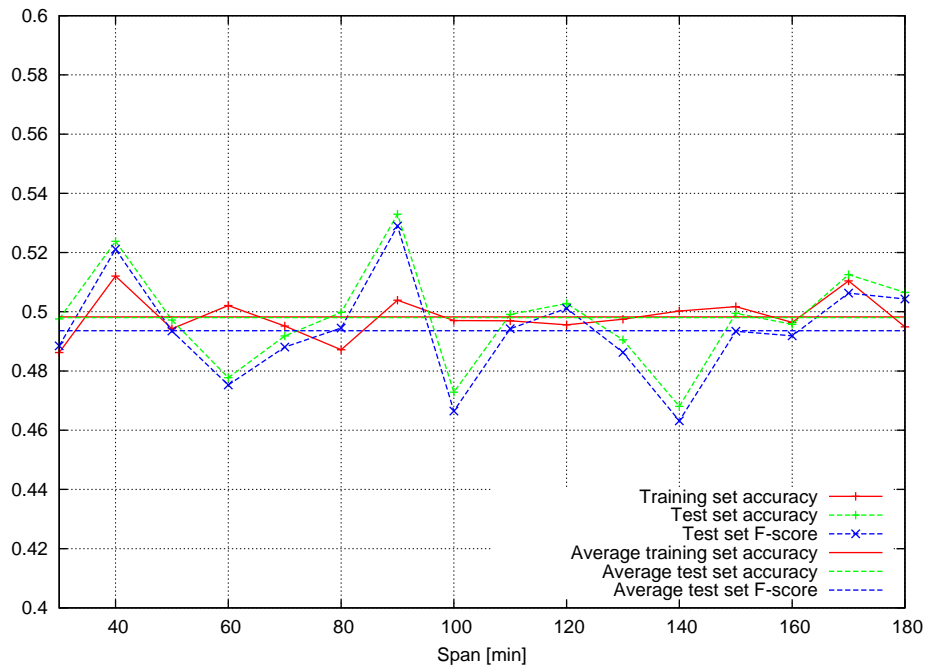


図 5.2: ランダム予測で予測期間を変化させた実験結果

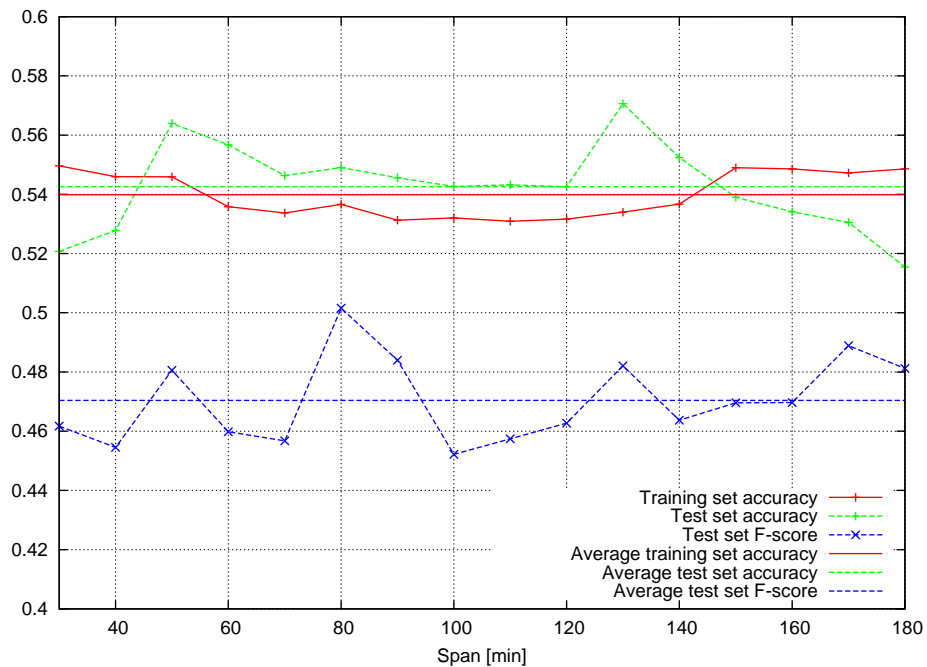


図 5.3: BoW のみで予測期間を変化させた実験結果

5.4 提案手法の個別適用実験

次に，以下の提案手法を個別に適用させた際の実験を行う．

タイトル 記事の本文ではなくタイトルを BoW にて特徴ベクトルに変換する

品詞フィルタ MeCab の品詞フィルタリングとして，表 4.1 のフィルタリングを行う

単語出現頻度フィルタ BoW にて特徴ベクトルに変換したあと，単語の出現頻度に基づく
フィルタリングを行う

5.4.1 タイトルのみで予測

まず，記事タイトルを用いた手法を適用し，予測期間を変化させた場合の実験結果は図 5.4 のようになる．テストセット正答率についてはベースラインと比較してパターンが変化しているものの，値としては近いものである．しかしながら F 値に関しては，ベースラインと比較して全体的に良くなっている傾向があり，40 分の時に最大で 0.5337 となっている．

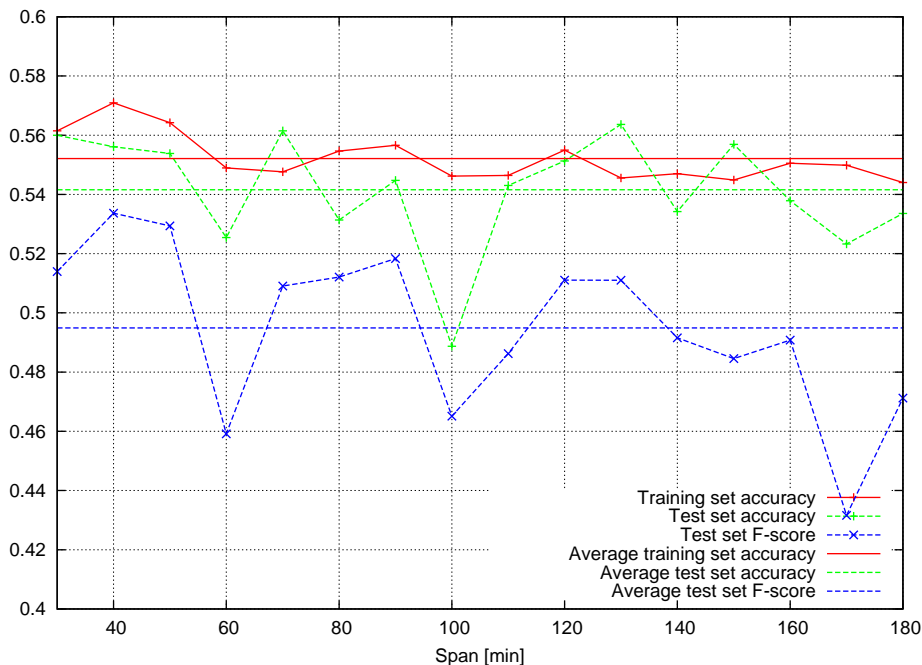


図 5.4: BoW+タイトルで予測期間を変化させた実験結果

5.4.2 品詞フィルタのみで予測

次に、品詞フィルタを適用し、予測期間を変化させた場合の実験結果は図 5.5 のようになる。結果はタイトルのみと同じように、テストセット正答率はベースラインと比べあまり大きく変わらないものの、F 値で改善が見られる。

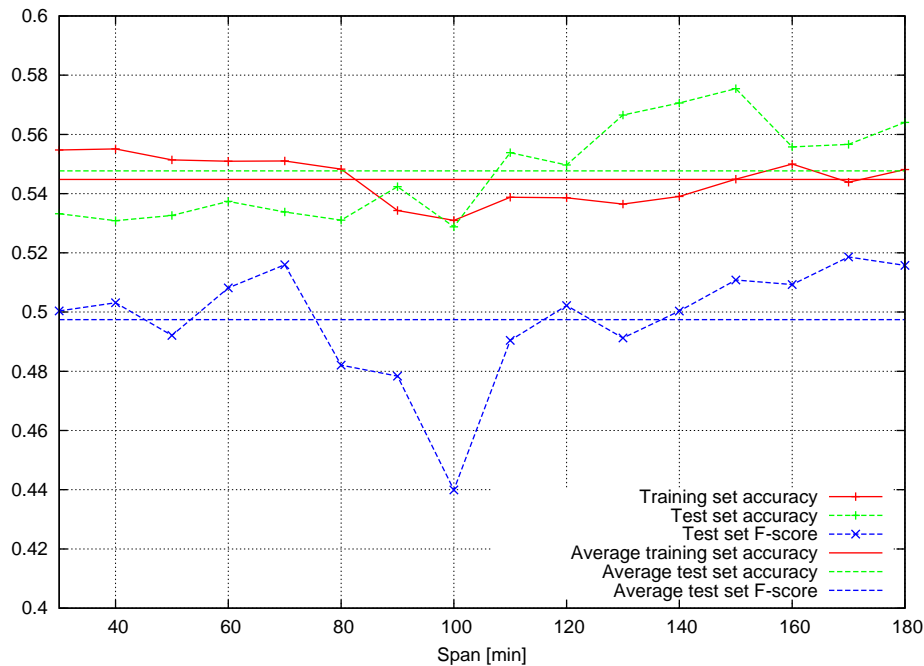


図 5.5: BoW+品詞フィルタで予測期間を変化させた実験結果

5.4.3 単語出現頻度フィルタのみで予測

最後に、単語出現頻度フィルタを適用し、単語出現頻度のしきい値である下限値と上限値を変化させた場合の実験結果は図 5.6, 5.7, 5.8 のようになる。この実験では、予測期間は 60 分で固定した。図 5.6 の訓練セット正答率を見ると、単語出現頻度の下限値と上限値は高い方が正答率が高かったが、図 5.7 のテストセット正答率については傾向を見ることはできない。しかしながら図 5.8 の F 値を見ると、下限値は 0.05 より高く、上限値は 0.75 より高い方が、結果は良かった。まとめると、単語出現頻度フィルタによって平均 F 値を改善させることができ、特に下限値の適用が有効だった。

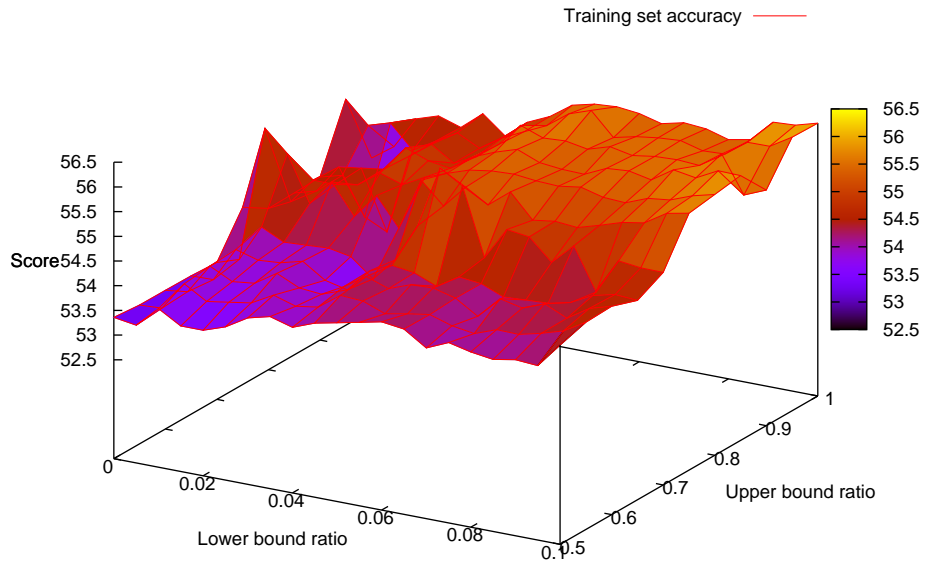


図 5.6: BoW+単語出現頻度で出現頻度のしきい値を変化させた実験の訓練セット正答率

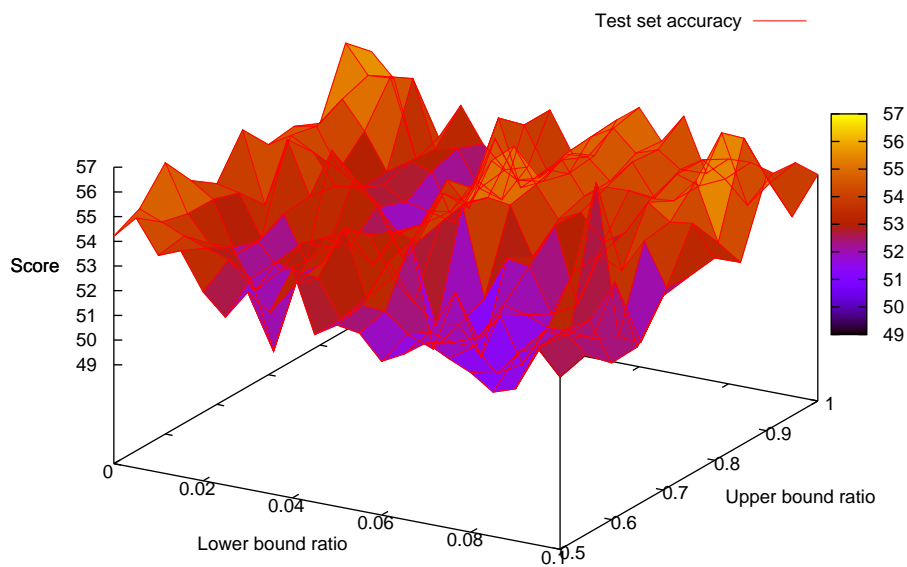


図 5.7: BoW+単語出現頻度で出現頻度のしきい値を変化させた実験のテストセット正答率

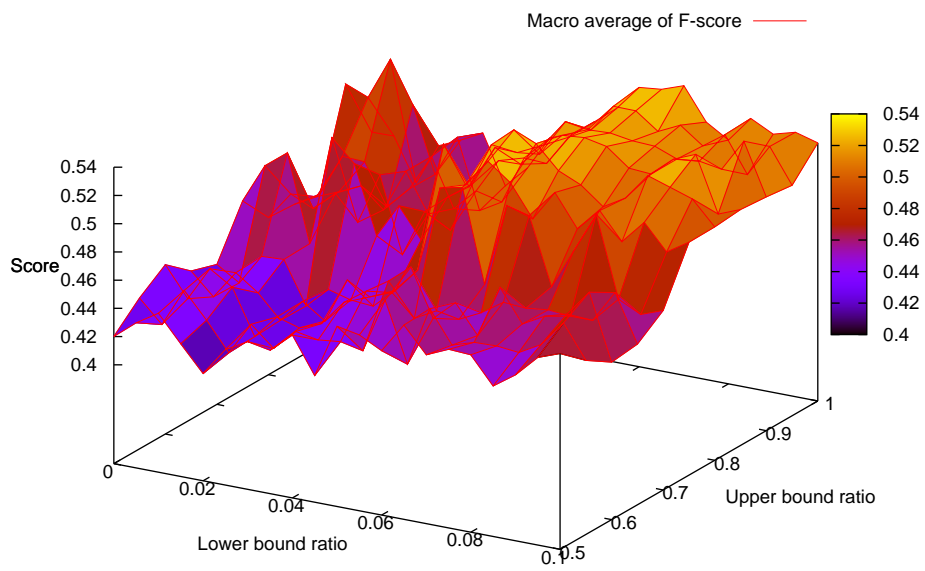


図 5.8: BoW+単語出現頻度で出現頻度のしきい値を変化させた実験の平均 F 値

5.5 提案手法の複数種組み合わせ実験

次に，前節で述べた3種類の提案手法を組み合わせ場合の実験結果を述べる．

5.5.1 タイトル+品詞フィルタで予測

まず初めに，タイトルと品詞フィルタを組み合わせた時に，予測期間を変化させた場合の実験結果は図5.9のようになる．それぞれの手法を単独で適用した時と比べ，テストセット正答率と平均F値は共に最高値と最低値に改善が見られる．これは，品詞フィルタを適用することによって特徴ベクトルのノイズが減り，結果が安定したためと考えられる．

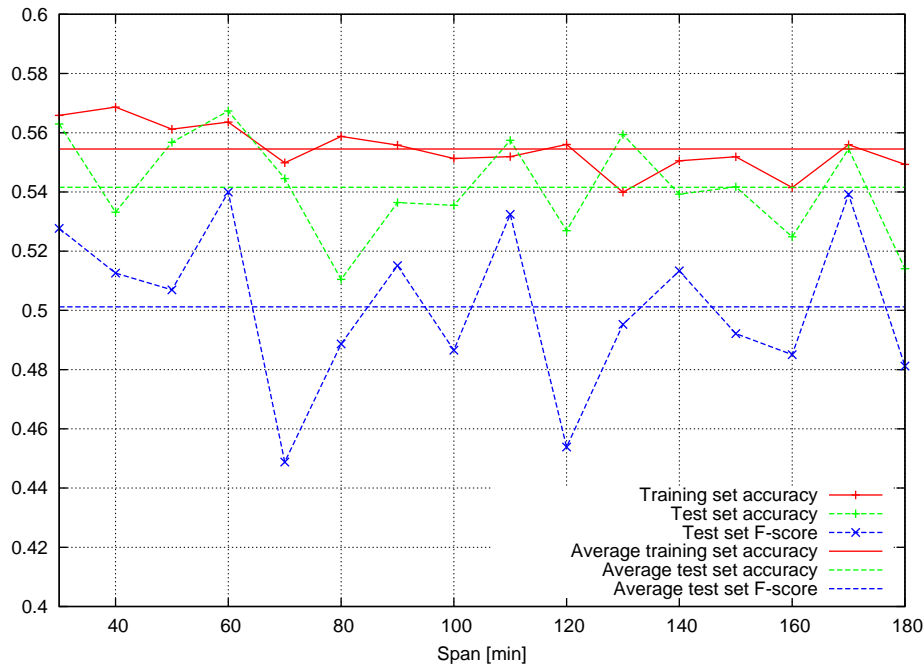


図 5.9: BoW+タイトル+品詞フィルタで予測期間を変化させた実験結果

5.5.2 品詞フィルタ+単語出現頻度フィルタで予測

次に，品詞フィルタと単語出現頻度フィルタを適用し，予測期間を変化させた場合の実験結果は図5.10のようになる．品詞フィルタのみを適用した場合と比べ，テストセット正答率については平均値が若干ながら改善が見られ，またF値に関しては大幅な改善が見られた．

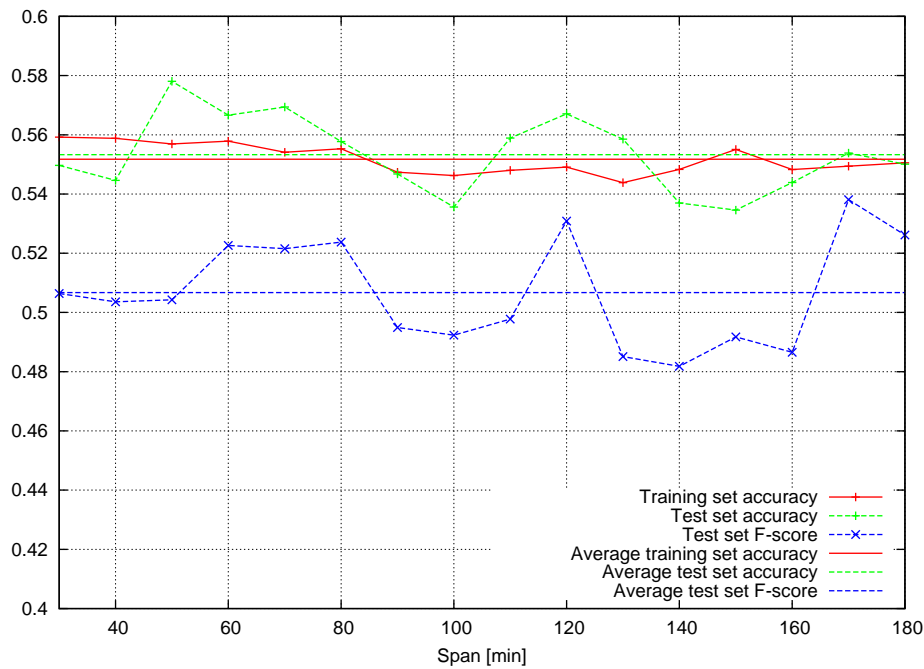


図 5.10: BoW+品詞フィルタ+単語出現頻度フィルタ [0.05, 1.0] で予測期間を変化させた実験結果

5.5.3 タイトル+単語出現頻度フィルタで予測

最後に、タイトルと単語出現頻度フィルタを適用し、予測期間を変化させた場合の実験結果は図 5.11 のようになる。タイトルのみを適用した場合と比べ、テストセット正答率とF値と共に平均値で改善が見られた。

5.6 全手法の適用実験

最後に、全手法を適用した場合の実験結果について述べる。

5.6.1 単語出現頻度のパラメータ実験

まず、単語出現頻度フィルタを単独で適用した実験では、適用対象が記事本文であったため、改めてタイトルと組み合わせた時の実験を行い、その結果は図 5.12, 5.13, 5.14 のようになった。

これらの結果の傾向は、単語出現頻度フィルタを単独で適用した場合と比べ大きく異なる。特に、記事本文で見られた上限値の調整による結果の差が非常に大きくなっている。これは

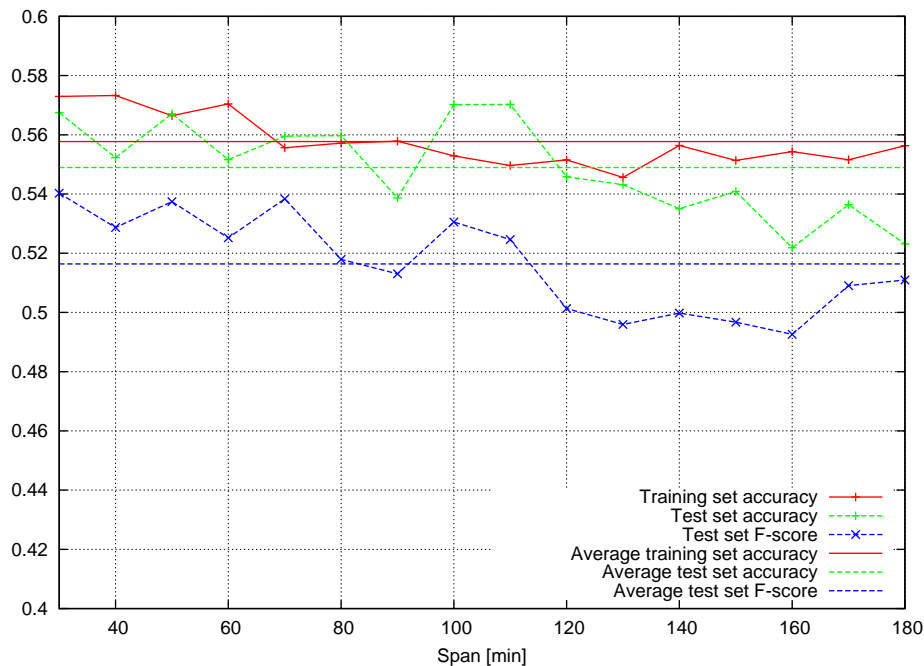


図 5.11: BoW+タイトル+単語出現頻度フィルタ [0.05, 1.0] で予測期間を変化させた実験結果

タイトルで使われる単語として、半分以上のタイトルで使われるような単語が非常に稀であるためと考えられる。ただし全体の傾向として、下限値と上限値ともに高めの方が結果が良い傾向は変わっていない。

5.6.2 予測期間のパラメータ実験

そして、単語出現頻度フィルタの下限値を 0.06, 上限値を 1.0 に設定し、予測期間を変化させた場合の実験結果は図 5.15 のようになる。なお、この実験では評価方法として仮想トレードのリターンも追加した。結果として、テストセット正答率の平均値が 56.55, F 値の平均が 0.5489 といずれの結果よりも大幅な改善が見られ、全ての実験を通して最も結果が良かった。

また仮想トレードのリターンは最低 3.5%, 最高 7.6%, 平均 5.1% という結果が得られた。これは、直接の比較対象としては適切ではないが、同期間の日経平均のリターン 4.0% よりも良かった。傾向としてはテストセット正答率と似た動きであるが、予測期間が 160 分以上の時に反してリターンが増加している。これは予測期間が長ければ長いほど、株価の値動きの幅(ボラリティ)が大きくなるので、1 回あたりのリターンが大きいためと考えられる。

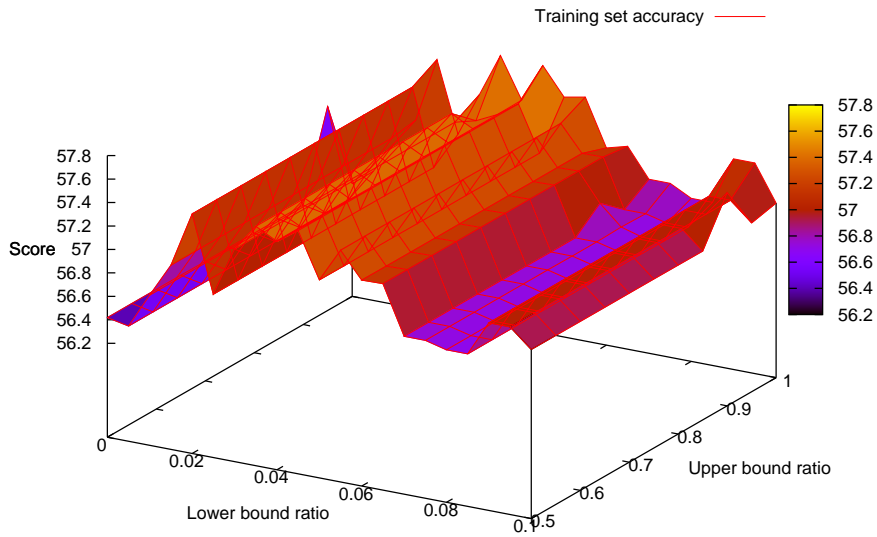


図 5.12: BoW+タイトル+品詞フィルタ+単語出現頻度で出現頻度のしきい値を変化させた実験の訓練セット正答率

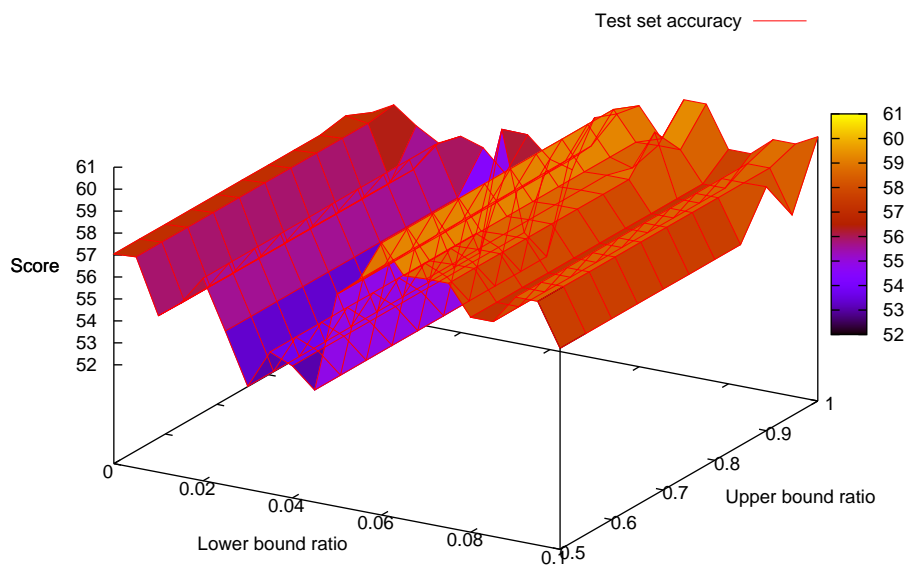


図 5.13: BoW+タイトル+品詞フィルタ+単語出現頻度で出現頻度のしきい値を変化させた実験のテストセット正答率

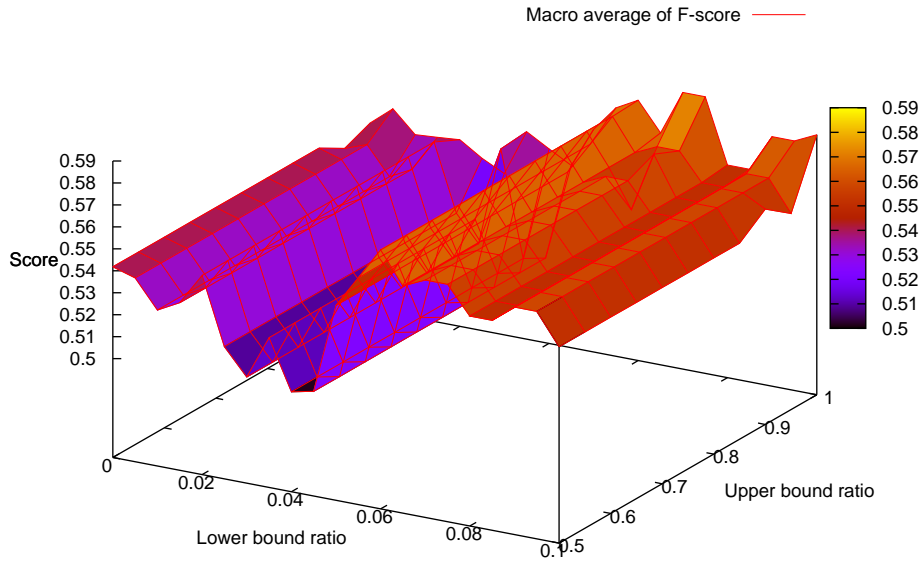


図 5.14: BoW+タイトル+品詞フィルタ+単語出現頻度で出現頻度のしきい値を変化させた実験の平均 F 値

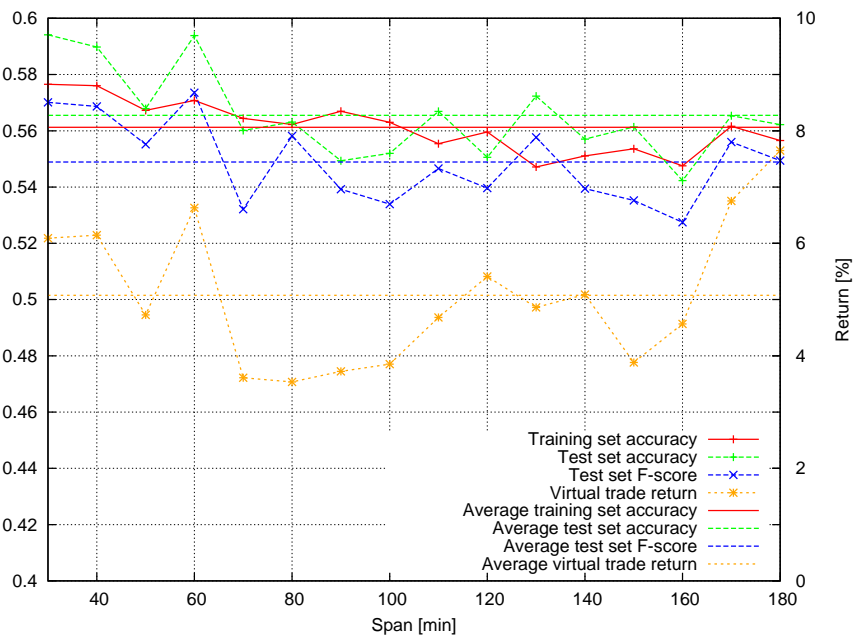


図 5.15: BoW+タイトル+品詞フィルタ+出現頻度フィルタ [0.06, 1.0] で予測期間を変化させた実験結果

5.7 最優秀結果の詳細

前節の実験の結果，予測期間を30分，単語出現頻度フィルタの下限値と上限値をそれぞれ0.06と1.0としたときに，評価値が最も良かった。³ この結果について，銘柄ごとの評価値を列挙したものが表5.2になる。

表5.2: 予測期間=30分, 単語出現頻度フィルタの下限値=0.06, 上限値=1.0の時の銘柄別の評価値

銘柄	訓練セット 正答率	テストセット 正答率	平均F値	リターン
トヨタ自動車	62.8571%	66.67%*	0.6667	10.1%
ソニー	57.6271%	64.29%*	0.5906	8.9%
ニコン	57.6014%	58.62%	0.5842	10.1%
ファナック	56.1576%	53.33%	0.5328	6.3%
ブリヂストン	58.9316%	58.93%	0.5892	7.6%
丸紅	57.3308%	58.14%	0.5700	6.0%
東京ガス	55.9454%	71.05%*	0.7054	3.9%
東京電力	54.8327%	64.44%	0.5853	3.8%
JX 日鉱日石エネルギー	56.5385%	58.70%	0.5705	2.0%
日清製粉グループ	61.2737%	48.28%	0.4727	-1.5%
鹿島建設	55.1515%	63.89%	0.6247	8.3%
三菱UFJファイナンシャル・グループ	61.39%	70.21%*	0.6278	9.8%
東日本旅客鉄道	55.2365%	53.70%	0.5291	5.3%
セブン&アイ・ホールディングス	54.9834%	50.00%	0.4246	5.9%
東京ドーム	58.9286%	50.94%	0.4780	4.9%
平均	57.65%	59.41%	0.5701	5.1%

*: 有意水準5%の二項検定で有意

テストセット正答率が有意水準5%の二項検定で有意だったのは，トヨタ，ソニー，東京ガス，三菱UFJの4社のみであったが，それ以外にも比較的正答率が高い銘柄が多い。F値に関しては，テストセット正答率と同じような水準で，正答率が高いがF値が低いような極端な銘柄は見られない。リターンに関してはボラリティに依存するので銘柄間で直接比較するのは難しいが，マイナスになっているのはテストセット正答率が唯一50%を下回った日清

³ 予測期間が60分の結果も同等に良く，僅かながら平均F値は60分の方が良かったが，全体的に時間が長くなればなるほど結果は悪くなっているため，ここでは30分の結果について詳細を述べる

製粉だけであった。

全体としては、「輸出入企業は為替相場に影響を受けるが、内需企業は影響を受けない」という仮説に沿った結果は得られなかった。これは仮説が部分的に間違っていることを意味し、現在の多様な企業形態においては為替相場の影響は業種によって一定でないと言える。つまり提案手法が有効であるかどうかは、業種別ではなく銘柄別に調べる必要がある。

5.8 日経平均223銘柄での実験

これまでの実験は、日経平均を構成する銘柄の中から15銘柄を抽出して実験を行い、結果から手法の有効性を確認できた。しかしながら15銘柄では銘柄に依存した影響を少なからず受ける可能性がある。そこで最も結果が良かった提案手法の組み合わせを用いて、日経平均を構成する225銘柄(2012年10月30日時点)を対象に、同じ実験を行うことで、15銘柄による実験結果の妥当性を検証する。

この実験の対象となった詳細な銘柄は、付録Cに掲載する。ただし証券コード3893の(株)日本製紙グループ本社と8815の東急不動産(株)は、実験データの期間中に吸収合併などにより上場廃止となったため、この2銘柄を除いた223銘柄で実験を行った。

結果は図5.16のようになった。結果としては、図5.15と傾向は似たものとなったが、より予測期間の違いによるゆらぎが小さくなった。この結果、15銘柄での実験は223銘柄による実験と差異が少ないことと、予測期間が長くなればなるほど各評価値が小さくなっていると見ることができる。また仮想トレードの結果は、15銘柄の時と比べリターンが高くなっているが、リターンの絶対値は銘柄のボラリティに依存するので、ここではその値が5%以上なため十分に高いと言える。

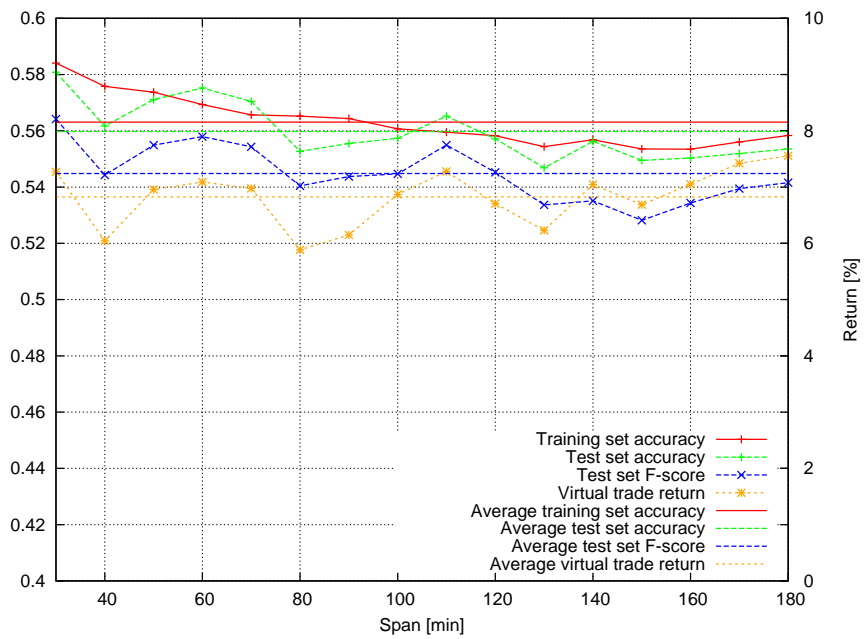


図 5.16: BoW+タイトル+品詞フィルタ+出現頻度フィルタ [0.06, 1.0] で予測期間を変化させた実験結果 (日経平均 223 銘柄)

第6章

考察

本章では、前章の株価動向予測実験に対する考察を行い、提案手法に関する評価を述べる。

6.1 為替ニュース

テキストマイニングするデータソースとして、近年盛んに行われている Twitter などに代表されるマイクロブログ [16, 17, 18] や、大衆紙などを対象とすることが多く、この選択肢の豊富さが金融テキストマイニングの研究を難しくさせる。

また対象が豊富なためベンチマークとなるデータが存在せず、自ずと各研究で使うデータが異なるため、同じデータによる実験を通じた純粋な手法の比較も難しいことが指摘できる。

本研究でもこの異なるデータであることを課題としつつも、既存研究とは違う着目点として、以下の点に主眼を置いた。

- 投資家らが投資情報を得るためのテキストであること
- 速報性を重視するため、ウェブ上であること
- テキストと株価の関係が明白であること

これらの点を考慮した結果、本研究ではウェブ上で証券会社が配信する為替ニュースを用いた。その結果、単純な BoW のみでもテストセット正答率は平均 54% を達成することができた。これは為替ニュースによる予測可能性を示すものであると考えられる。

ただし課題としては、一日の為替ニュースの中で極端に為替相場が変動しない限り、内容に差異が出にくいいため似た予測をする傾向にあり、一日のトレンドが安定していない場合は予測を外す傾向にある。そのため同じ日の複数のニュースの連続性を考える必要がある。

6.2 タイトルの利用

既存研究では、ニュースをテキストマイニングの対象とするとき、自然に記事本文を利用している。これはひとえに記事本文がそのニュースの本質であり、情報量が大きいためである。

ところが実際のニュース本文には、そのニュースの本質とは離れたテキストが含まれていることが多い。例えば本質の前提となるようなサブトピックスなどが挙げられる。また署名や注釈といった、ノイズのようなテキストも往々にして含まれる。このことは、いくら情報量が多いからといえども、記事本文から特徴量を抽出するうえで不都合である。

またヘッドラインニュースのように、速報としてタイトルだけ先に配信し、あとから内容

を再送する場合もよく見られる。このことから、投資家はタイトルのみ反応して取引を行っている可能性があると言える。つまりタイトルが最も重要であり、内容を吟味するのは二の次であると言える。

そこで本研究では、タイトルから特徴量を抽出すること提案した。これは上記のように投資家の目線としてと、そのニュースの本質のみがタイトルに含まれていると考えたからである。

実験結果では、タイトルのみを単独で適用した場合や他の手法と組み合わせ場合、テストセット正答率や平均F値において結果の改善を確認することができた。これはタイトルに情報密度の高さがあると言える。しかしながら絶対的な情報量としては記事本文にあるので、今後の研究としては記事本文からタイトルと同等の情報密度に圧縮しつつ、情報量を失わないような手法の研究が要請される。

6.3 品詞フィルタ

本研究では、アドホックな品詞フィルタを定義したが、内容としては細かいものであり、またその内容を検証できるような実験をするに至っていない。これは品詞フィルタの品詞による結果の差を調べるのが、本研究の大きな目的ではないためである。

しかしながら、アドホックに定義したフィルタでありながら、その有効性を実験結果から確認することができた。このことは、今後の品詞フィルタの更なる改良の余地が存在することを意味する。

また本研究では形態素解析器として MeCab+IPA 辞書を用いたため、それに準じた品詞フィルタとなっている。これは形態素解析器非依存、ひいては言語非依存な品詞フィルタを定義するうえで、望ましくないことである。そのため、今後はまず形態素解析器に依存しない、日本語の品詞によるフィルタリングを考える必要がある。言語非依存な手法については、言語間の品詞の差が少なからず存在することを考えると、十分に検証したうえでその可能性を検討する必要がある。

6.4 単語出現頻度フィルタ

本研究で提案した単語出現頻度フィルタは、従来のストップワード除去をカバーする手法であり、また言語非依存な手法を目的とした手法である。

このフィルタを単独で適用した実験の結果からはっきりとした傾向を掴むことは難しいが、

平均 F 値の面から見れば下限値を 0.05 以上 0.1 以下，上限値を 0.75 以上 1.0 以下の組み合わせの場合が，他の組み合わせと比べ結果が良かった．また品詞フィルタと組み合わせ実験結果では，ストップワードが除去されてるために上限値による差が非常に少ないが，上限値が 1.0 の場合，結果が良くなっている傾向が見れる．

このことは，単語出現頻度フィルタの上限値によるストップワード除去の効果よりも，必要な特徴量までもが失われている可能性を示唆する．つまりストップワード除去を目的とした上限値の設定は逆効果で，もしストップワードを除去したいのならば単語フィルタなり品詞フィルタなりで対応しなければいけないと言える．

ただし，下限値の設定による結果の改善はうかがえるので，単語出現頻度フィルタ自体の有効性が無くなったわけではない．この下限値としては，今回 0.06 という値が最も良いと判断したが，用いる言語やテキストによって最も良い値が異なることは容易に想像がつくので，違う研究に用いる際は改めてこのパラメータの調整を行う必要がある．そのためパラメータ探索が要求されるという点に関しては，この手法のデメリットと言える．

6.5 平均 F 値

既存研究では，評価の尺度として主に精度を用いてきた．しかしながら精度だけで評価を論ずることは，明らかに尺度として足りない．なぜなら，精度は正例と負例の正答を同じ尺度で評価するため，精度だけでは正例と負例の予測の偏りを評価することができない．

そこで本研究では，この足りない尺度である偏りに対応した指標として F 値，特に平均 F 値を評価に用いた．

全ての実験を通して，平均 F 値がテストセット正答率と似た動きであることを確認できる．この平均 F 値を計算する過程で精度の指標も使っているためである．しかしながら正答率が近い値の時に結果の優劣を決める次点の指標として，平均 F 値が高いほうがより良い結果であることは，その計算方法から自明である．

よって今後の同様の研究では，この平均 F 値による評価もなされることを期待したい．

6.6 予測期間

既存研究では予測期間を 1 日とする研究が多い中，あえて 30 分から 180 分の短期予測を実験した．これは予測期間が長期になればなるほど，ニュース以外の要因による株価変動が起こり，予測精度が下がってしまうと考えたためである．

15 銘柄での実験では結果の変動が激しく確認しづらいが、最後の 223 銘柄での実験では、テストセット正答率と平均 F 値において予測期間に対する緩やかな下落傾向を確認できる。これはすなわち、上記の仮説を支持するものである。

そうならば 30 分未満の方が更に予測精度は高くなるはずであるが、時間が短くなればなるほどニュースが投資家らに浸透しないがために、十分に株価に反映されるなくなる。そのため 30 分未満でもピークの場所が存在することになる。

実際に既存研究で、このピークの時間を 20 分とする研究 [19] がある。本研究でも、この 20 分という時間を含め前後 5 分程度の時間で実験することが望ましかったが、使用したニュースデータの中に 8 時半のものがああり、データ数の確保という課題があったという都合上、20 分という時間を試すことができなかったことは今後の改善点である。

第7章

結論

7.1 まとめ

本研究では、投資家を対象としたウェブ上の為替ニュースを SVM によって機械学習し、株価動向を予測するための手法として、為替市況に関するニュースに制限する手法、ニュースの記事本文ではなくタイトルを用いる手法、MeCab によって形態素解析する際の品詞フィルタ、そして全体の単語出現頻度に基づくフィルタの4つの手法を提案し、それぞれの手法を個別もしくは組み合わせて評価した。

実験結果としては、それぞれの手法において正答率や F 値の増加、結果の安定性の向上を確認できたので、手法の有効性を確認できた。そして全手法を組み合わせた時に、テストセット正答率と F 値はそれぞれ最大 58.08% と 0.5642、平均でも 55.97% と 0.5448 となり、共に高い値となり、仮想トレードのリターンの観点からもその有効性を確認することができた。

学習するニュースとして為替市況に関するものだけに制限する手法に関しては、ベースラインの実験だけでも正答率を平均 54.25% に押し上げることができた。これは、そもそも株価が為替だけに左右されるわけではないことなどを考えると、その有効性は高いと考えられる。

品詞フィルタに関しては、独断と偏見で除外する品詞を決定したので、さらにどの品詞を除外するか、また逆に除外対象から外すかで更なる手法の改善に繋がる可能性がある。また決定方法自体がアドホックなので、決定方法に関して統一的な枠組みが必要とされる。

単語出現頻度フィルタに関しては、実験結果からは下限値を 0.06 より高く、上限値は 1.0 のままとする方が良いことを確認できたが、この閾値はデータソースによって変化する可能性が高いため、他のデータソースに単語出現頻度フィルタを適用する場合はその都度調整をする必要があると考えられる。

また本研究では予測期間による差を見るのが主目的ではなく、15 銘柄による実験ではその有意な差を見るができなかったが、223 銘柄による実験で僅かながら予測期間に対する精度の下落傾向を確認することができた。これは断言することはできないが、予測期間が長くなればなるほど他の要因による株価への影響が大きくなるためであると考えられる。今後は 30 分未満の短期間の予測と 180 分以上の長期間の予測を実験して、より詳細に検討する必要がある。

7.2 今後の展望

この研究の今後の展望としては、前章でも課題として若干触れたが、大きく以下の点が考えられる。

- 為替相場と株価の相関性が結果と一致するか
単純な「輸出入企業は為替相場の影響を受け、内需企業は影響を受けない」という仮定と矛盾する実験結果が得られたことを受け、自然に考えるのならば、為替相場の影響の受け方は業種によって一定ではないと考えられるので、これを検証する必要がある。
- 記事本文の自動要約
本研究では、ノイズが少ないという理由でニュースのタイトルを特徴量に変換したが、記事本文と比較して絶対的な情報量が少ないという点は否定できない。そのため、将来的には記事本文を特徴量として利用するために、いかにノイズとなりえる文章を排除するかが課題となる。そこで近年注目されている、文章の自動要約技術 [20] を活用することで、本文の情報量をなるべく減らさずに、ノイズを除去することが可能となる。
- 学習データの拡充
本研究で用いた学習データは、訓練セットで最大 631 個、テストセットで最大 60 個とビッグデータと言うには余りにも少ない。これは、用いたニュースが比較的新しいサービスであることと、そのサービスが定時制で 1 日に 4 本しか配信されないことに起因する。そこでもっと大規模な学習データで学習するために、あらゆるニュースの中から為替市況に関するニュースのみに絞る手法が要請される。
- 分類クラスの拡張
本研究では、分類クラスとして用いたのは株価の上昇・下降という 2 クラスだった。しかし現実には、多少の値動きはあるが、明確なトレンドがあるわけではないような状況や、強い上昇・下降トレンドなどといった、単純な上昇・下降の 2 クラスでは片付けられない。実際の投資家の心理としては、強いトレンドの発生が予測される場合のみに売買を行いたいので、この要求に沿うような手法が望まれる。
- 特徴量抽出手法の改良
本研究では特徴量抽出に関する手法に拘らず、単純な Bag-of-Words による特徴量抽出を採用した。しかし既存研究の中には、BoW 以外の特徴量の抽出手法や選択手法を用いている例 [21] もあり、それぞれ一定の成果を上げている。そこで更なる精度の向上のために、特徴量に関する手法の改善が望まれる。

謝辞

まず始めに、この研究は多くの人の支えによって成り立っていることをここに深く感謝申し上げます。

この研究に対して一番良くアドバイスをして頂いた、Danushka 先生には感謝の念が尽きません。その幅広い知見から、研究について様々な提案をして頂き、また結果の評価についても手助けをして頂きました。途中からイギリスの大学に転属され、東大の籍では無くなったにも関わらず、メール等を通じて支援して頂いたことを、深く深くお礼申し上げます。

また自然言語処理に関しては専門ではないにも関わらず、研究内容から発表まで幅広くアドバイスをして頂いた伊庭先生にも、深く感謝申し上げます。

既にご卒業されていますが、渡辺晃生さんには研究室のサーバー管理方法について伝授して頂きました。そのため小規模ネットワークを持つサーバーに関するスキルを身につけることができました。この貴重な機会を作って頂いた晃生さんに感謝致します。

他研究室の博士・修士の先輩方や同期、また学部の後輩とは、他愛の無い雑談によって楽しい研究生を送ることができました。有難うございます。

学校外では、研究に失敗して辛く落ち込んでいる時でも精神的に支えてくれた彼女なしでは、この研究を成し遂げることはできませんでした。研究以外の楽しみを一緒にすることができて、大変嬉しかったです。有難うございます。これからもよろしくお願ひします。

最後に、修士まで何一つ不自由なく勉強・研究できるよう絶え間なく経済的援助をしてくれた父親に、心からお礼申し上げます。いつかは親孝行します。

参考文献

- [1] 藤原茂章. 最近の株価と為替の同時相関関係の強まりについて. Technical report, 日本銀行金融市場局, December 2013.
- [2] Gabriel Pui Cheong Fung, J.X. Yu, and Wai Lam. Stock prediction: Integrating text mining approach using real-time news. In *Computational Intelligence for Financial Engineering, 2003. Proceedings. 2003 IEEE International Conference on*, pp. 395–402, 2003.
- [3] Eugene F Fama. Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, Vol. 25, No. 2, pp. 383–417, May 1970.
- [4] V. Vapnik. *Statistical Learning Theory*. Wiley, Chichester, GB, 1998.
- [5] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1986.
- [6] 工藤拓. Mecab: Yet another part-of-speech and morphological analyzer. <http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html>.
- [7] 工藤拓, 山本薫, 松本裕治. Conditional random fields を用いた日本語形態素解析 (解析). 情報処理学会研究報告. 自然言語処理研究会報告, Vol. 2004, No. 47, pp. 89–96, may 2004.
- [8] 藏本貴久, 和泉潔, 吉村忍, 石田智也, 中嶋啓浩, 松井藤五郎, 吉田稔, 中川裕志. 新聞記事のテキストマイニングによる長期市場動向の分析. 人工知能学会論文誌, Vol. 28, No. 3, pp. 291–296, 2013.
- [9] 辻洋平, 古宮嘉那子, 小谷善行. Web ニュース中の複数企業に対応した株価予測. 電子情報通信学会技術研究報告. IBISML, 情報論的学習理論と機械学習, Vol. 110, No. 476, pp. 109–113, mar 2011.

- [10] Robert P. Schumaker and Hsinchun Chen. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Trans. Inf. Syst.*, Vol. 27, No. 2, pp. 12:1–12:19, March 2009.
- [11] Robert P. Schumaker and Hsinchun Chen. A discrete stock price prediction engine based on financial news. *Computer*, Vol. 43, No. 1, pp. 51–56, January 2010.
- [12] Robert P. Schumaker, Yulei Zhang, Chun-Neng Huang, and Hsinchun Chen. Evaluating sentiment in financial news articles. *Decis. Support Syst.*, Vol. 53, No. 3, pp. 458–464, June 2012.
- [13] Chih-Chung Chang and Chih-Jen Lin. LIBSVM. <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>.
- [14] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, Vol. 2, pp. 27:1–27:27, 2011. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [15] Chih-Chung Chang and Chih-Jen Lin. LIBSVM FAQ. <http://www.csie.ntu.edu.tw/~cjlin/libsvm/faq.html>.
- [16] Johan Bollen, Huina Mao, and Xiao-Jun Zeng. Twitter mood predicts the stock market. *Computer*, Vol. 2, No. 1, pp. 1–8, 2010.
- [17] J Bollen. Twitter mood as a stock market predictor. *Computer*, No. October, pp. 91–94, 2011.
- [18] Salah Bouktif and Mamoun Adel Awad. Ant colony based approach to predict stock market movement from mood collected on twitter. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM '13*, pp. 837–845, New York, NY, USA, 2013. ACM.
- [19] Gyozo Gidofalvi. Using news articles to predict stock price movements, 2001.
- [20] 奥村学, 難波英嗣. テキスト自動要約. オーム社, 2005.
- [21] Michael Hagenau, Michael Liebmann, Markus Hedwig, and Dirk Neumann. Automated news reading: Stock price prediction based on financial news using context-specific features.

- 2013 46th Hawaii International Conference on System Sciences, Vol. 0, pp. 1040–1049, 2012.
- [22] 丸山健, 梅原英一, 諏訪博彦, 太田敏澄. インターネット株式掲示板の投稿内容と株式市場の関連性. 証券アナリストジャーナル, 第46巻第11・12号, pp. 110–127, 2008.
- [23] Blake LeBaron, W.Brian Arthur, and Richard Palmer. Time series properties of an artificial stock market. *Journal of Economic Dynamics and Control*, Vol. 23, No. 9-10, pp. 1487 – 1516, 1999.
- [24] Gabriel Pui Cheong Fung, Jeffrey Xu Yu, and Wai Lam. News sensitive stock trend prediction. In *Proceedings of the 6th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, PAKDD '02, pp. 481–493, London, UK, UK, 2002. Springer-Verlag.
- [25] Youngohc Yoon and G. Swales. Predicting stock price performance: a neural network approach. *System Sciences*, Vol. 4, pp. 156–162, 1991.
- [26] Gangshu Cai and Peter R. Wurman. Monte carlo approximation in incomplete information, sequential auction games. *Decis. Support Syst.*, Vol. 39, No. 2, pp. 153–168, April 2005.
- [27] 石井健一郎, 前田英作, 上田修功, 村瀬洋. わかりやすいパターン認識. オーム社.
- [28] 和泉潔, 松井藤五郎. 金融テキストマイニング研究の紹介. 情報処理学会誌, Vol. 53, No. 9, pp. 932–937, 2012.
- [29] 和泉潔, 後藤卓, 松井藤五郎. 経済テキスト情報を用いた長期的な市場動向推定. 情報処理学会誌, Vol. 52, No. 12, pp. 3309–3315, 2011.
- [30] 和泉潔, 後藤卓, 松井藤五郎. テキスト情報による金融市場変動の要因分析. 人工知能学会論文誌, Vol. 25, No. 3, pp. 383–387, 2010.
- [31] Xiangyu Tang, Chunyu Yang, and Jie Zhou. Stock price forecasting by combining news mining and time series analysis. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '09, pp. 279–282, Washington, DC, USA, 2009. IEEE Computer Society.

- [32] Wei Fan and Toyohide Watanabe. Dynamic prediction of forthcoming trends in stock prices from news articles. In *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics, WIMS '12*, pp. 16:1–16:9, New York, NY, USA, 2012. ACM.
- [33] Michael Hagenau, Michael Liebmann, Markus Hedwig, and Dirk Neumann. Automated news reading: Stock price prediction based on financial news using context-specific features. In *Proceedings of the 2012 45th Hawaii International Conference on System Sciences, HICSS '12*, pp. 1040–1049, Washington, DC, USA, 2012. IEEE Computer Society.
- [34] Ann Devitt and Khurshid Ahmad. Sentiment polarity identification in financial news: A cohesion-based approach. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 984–991, Prague, Czech Republic, June 2007. Association for Computational Linguistics.
- [35] Paul C. Tetlock. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, Vol. 62, No. 3, pp. 1139–1168, 2007.
- [36] 岡田克彦, 中元政一, 東高宏, 羽室行信. 負け犬は誰だ? 証券アナリストの格下げにより価値を失う企業の特徴について, 2011.
- [37] Anshul Mittal and Arpit Goel. Stock prediction using twitter sentiment analysis. Technical report, Stanford University, 2011.
- [38] 奥村学, 高村大也. 言語処理のための機械学習入門. コロナ社.
- [39] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, Vol. 9, pp. 1871–1874, 2008.
- [40] ビショップ CM. パターン認識と機械学習下: ベイズ理論による統計的予測. 丸善出版, 2008.
- [41] G. Gidófalvi and C. Elkan. Using news articles to predict stock price movements. Technical report, Department of Computer Science and Engineering, University of California, 2003.
- [42] Zellig Harris. Distributional structure. *Word*, Vol. 10, No. 23, pp. 146–162, 1954.

発表文献

- [1] 石黒 祐輔, ダヌシカ・ボレガラ, 伊庭 斉志.
為替ニュース記事を用いた *SVM* による株価予測
第 12 回 人工知能学会 金融情報学研究会.

付録

A MeCab における IPA 辞書の品詞体系

IPA 辞書はもともと MeCab より前に開発された ChaSen という形態素解析器のために構築された辞書である。MeCab ではそれに対して新しい要素を加えたものを利用している。名前に IPA が含まれているため独立行政法人 情報処理推進機構 (IPA) が管理しているように感じられるが、直接は関係ない。

表 A.1: MeCab における IPA 辞書の品詞体系

品詞	細分類 1	細分類 2	細分類 3	例
その他	間投	*	*	よ, ア
フィラー	*	*	*	ええと, あの, その
感動詞	*	*	*	すいません
記号	アルファベット	*	*	Abc
	一般	*	*	!?
	括弧開	*	*	< 「
	括弧閉	*	*	> 」
	句点	*	*	.
	空白	*	*	
形容詞	読点	*	*	,
	自立	*	*	古い
	接尾	*	*	ぼい
	非自立	*	*	にくい

表 A.1: MeCab における IPA 辞書の品詞体系

品詞	細分類 1	細分類 2	細分類 3	例
助詞	格助詞	一般	*	で, にて
		引用	*	と
		連語	*	とかいう
	係助詞	*	*	こそ
	終助詞	*	*	ね, な
	接続助詞	*	*	けれども
	特殊	*	*	かな
	副詞化	*	*	に, と
	副助詞	*	*	など, ぐらい
	副助詞 / 並立助詞 / 終助詞	*	*	か
	並立助詞	*	*	と
連体化	*	*	の	
助動詞	*	*	*	です, ます
接続詞	*	*	*	けれど
接頭詞	形容詞接続	*	*	超, くそ
	数接続	*	*	約
	動詞接続	*	*	相, 御
	名詞接続	*	*	不, 非
動詞	自立	*	*	書く
	接尾	*	*	させる
	非自立	*	*	くださる
副詞	一般	*	*	おどおど
	助詞類接続	*	*	全く

表 A.1: MeCab における IPA 辞書の品詞体系

品詞	細分類 1	細分類 2	細分類 3	例	
名詞	サ変接続	*	*	増加	
	ナイ形容詞語幹	*	*	さりげ	
	一般	*	*	コスト	
	引用文字列	*	*	いわく	
	形容動詞語幹	*	*	自動的	
	固有名詞	一般	*	*	富士山
		人名	一般	*	豊臣秀吉
			姓	*	佐藤, 田中
			名	*	太郎, 次郎
		組織	*	*	情報処理学会
	地域	一般	*	東京	
			国	*	日本, アメリカ
	数	*	*	0, 1, 99	
	接続詞的	*	*	兼	
	接尾	サ変接続	*	*	化
		一般	*	*	汁
		形容動詞語幹	*	*	好き
		助数詞	*	*	キ口
		助動詞語幹	*	*	そう
		人名	*	*	様
		地域	*	*	行き
		特殊	*	*	っぷり
	副詞可能	*	*	明日	
	代名詞	一般	*	*	君
		縮約	*	*	そりゃあ
	動詞非自立的	*	*	ちょうだい	
	特殊	助動詞語幹	*	*	そう
	非自立	一般	*	*	はず
		形容動詞語幹	*	*	みたい
助動詞語幹		*	*	そう, よう	
副詞可能		*	*	以下	
副詞可能	*	*	全て		
連体詞	*	*	*	そういう	

B F値の求め方

表 B.1 に、2 クラス分類の結果である混同行列の示す。

表 B.1: 混同行列

予測 \ 実際	正例	負例
正例	TP	FP
負例	FN	TN

この時の (正例の)F 値¹ の求め方は、以下の通りである。

$$\text{精度 (Precision)} = \frac{TP}{TP + FP} \quad (\text{B.1})$$

$$\text{再現率 (Recall)} = \frac{TP}{TP + FN} \quad (\text{B.2})$$

$$\text{F 値 (F-value)} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (\text{B.3})$$

¹文書分類のタスクでは負例の F 値という概念がないため、正例が自明である

C 日経平均 225 銘柄

本研究で実験対象とした，2012年10月30日時点の日経平均を構成する225銘柄を表C.1に掲載する．

表 C.1: 日経平均 225 銘柄 (2012年10月30日時点)

証券コード	企業名
4151	協和発酵キリン (株)
4502	武田薬品工業 (株)
4503	アステラス製薬 (株)
4506	大日本住友製薬 (株)
4507	塩野義製薬 (株)
4519	中外製薬 (株)
4523	エーザイ (株)
4568	第一三共 (株)
6479	ミネベア (株)
6501	(株) 日立製作所
6502	(株) 東芝
6503	三菱電機 (株)
6504	富士電機 (株)
6506	(株) 安川電機
6508	(株) 明電舎
6674	(株) ジーエス・ユアサ コーポレーション
6701	日本電気 (株)
6702	富士通 (株)
6703	沖電気工業 (株)
6752	パナソニック (株)
6753	シャープ (株)
6758	ソニー (株)
6762	T D K (株)
6767	ミツミ電機 (株)
6770	アルプス電気 (株)
6773	パイオニア (株)
6841	横河電機 (株)
6857	(株) アドバンテスト
6902	(株) デンソー

表 C.1: 日経平均 225 銘柄 (2012 年 10 月 30 日時点)

証券コード	企業名
6952	カシオ計算機 (株)
6954	ファナック (株)
6971	京セラ (株)
6976	太陽誘電 (株)
7735	大日本スクリーン製造 (株)
7751	キヤノン (株)
7752	(株) リコー
8035	東京エレクトロン (株)
7201	日産自動車 (株)
7202	いすゞ自動車 (株)
7203	トヨタ自動車 (株)
7205	日野自動車 (株)
7211	三菱自動車工業 (株)
7261	マツダ (株)
7267	本田技研工業 (株)
7269	スズキ (株)
7270	富士重工業 (株)
4543	テルモ (株)
4902	コニカミノルタホールディングス (株)
7731	(株) ニコン
7733	オリンパス (株)
7762	シチズンホールディングス (株)
9412	(株) スカパー J S A T ホールディングス
9432	日本電信電話 (株)
9433	K D D I (株)
9437	(株) エヌ・ティ・ティ・ドコモ
9613	(株) エヌ・ティ・ティ・データ
9984	ソフトバンク (株)
8303	(株) 新生銀行
8304	(株) あおぞら銀行
8306	(株) 三菱 U F J フィナンシャル・グループ
8308	(株) りそなホールディングス
8309	三井住友トラスト・ホールディングス (株)
8316	(株) 三井住友フィナンシャルグループ

表 C.1: 日経平均 225 銘柄 (2012 年 10 月 30 日時点)

証券コード	企業名
8331	(株) 千葉銀行
8332	(株) 横浜銀行
8354	(株) ふくおかフィナンシャルグループ
8355	(株) 静岡銀行
8411	(株) みずほフィナンシャルグループ
8253	(株) クレディセゾン
8601	(株) 大和証券グループ本社
8604	野村ホールディングス (株)
8628	松井証券 (株)
8630	N K S J ホールディングス (株)
8725	M S & A D インシュアランスグループホールディングス (株)
8729	ソニーフィナンシャルホールディングス (株)
8750	第一生命保険 (株)
8766	東京海上ホールディングス (株)
8795	(株) T & D ホールディングス
1332	日本水産 (株)
1334	(株) マルハニチロホールディングス
2002	(株) 日清製粉グループ本社
2269	明治ホールディングス (株)
2282	日本ハム (株)
2501	サッポロホールディングス (株)
2502	アサヒグループホールディングス (株)
2503	キリンホールディングス (株)
2531	宝ホールディングス (株)
2801	キッコーマン (株)
2802	味の素 (株)
2871	(株) ニチレイ
2914	日本たばこ産業 (株)
3086	J . フロント リテイリング (株)
3099	(株) 三越伊勢丹ホールディングス
3382	(株) セブン & アイ ・ ホールディングス
8233	(株) 高島屋
8252	(株) 丸井グループ
8267	イオン (株)

表 C.1: 日経平均 225 銘柄 (2012 年 10 月 30 日時点)

証券コード	企業名
8270	ユニー (株)
9983	(株) ファーストリテイリング
4324	(株) 電通
4689	ヤフー (株)
4704	トレンドマイクロ (株)
9602	東宝 (株)
9681	(株) 東京ドーム
9735	セコム (株)
9766	コナミ (株)
1605	国際石油開発帝石 (株)
3101	東洋紡 (株)
3103	ユニチカ (株)
3105	日清紡ホールディングス (株)
3401	帝人 (株)
3402	東レ (株)
3861	王子ホールディングス (株)