概要

本研究は, 教師あり機械学習を用いることにより, 毎日新聞のテキストデータから日経平均の騰落の予測を行う. また, 機械学習が使用した素性を分析することで株式相場や経済に関わる知見を取得することを目指す. 機械学習の学習データには, 毎日新聞のテキストデータを用いる.

本研究の成果は 2 つある. 1 つ目は, 日経平均の予測の実験で, 毎日新聞の朝刊からその日の始値と終値の差の予測を行った. 機械学習の分類先を上昇・下降・変化なしの 3 分類としたとき, 正解率は一番高いもので 0.453 であり, ある程度の予測は行えていた. Buy&Hold 法をベースラインとして, 年間の平均利益と年間の赤字の最低値を比較すると, 平均利益は 7 割程度となっているが, 赤字の最低値は半分以下であった. 提案手法では, 利益に対して赤字の最低値が低くなっており, ベースライン手法より有用な点もあった.

2 つ目は, 機械学習に使った素性の分析を行ったことにより, 新聞記事から株式相場や経済に関わる知見を抽出できたことである.　まず, 毎日新聞の朝刊から 2 日前の終値と前日の終値の差の推定を行った. 次に機械学習に使った素性の株価上昇の正規化 α 値が上位のものと下位のものを調べ, それらを人手で考察し, また, 頻度の分析を行うことで株式相場や経済に関わる知見を取得することができた.

目 次

表 目 次

第1章　はじめに

近年では、インターネットや新聞などのテキストデータを数学やコンピュータを駆使して数値化し, 分析することによって, 値動きのある金融商品のリスクヘッジやリスクマネジメントに役立たせるといった研究が盛んに行われている.

先行研究 [1] では日経平均の上昇下降の予測は行っているが, 経済に関わる知見の獲得は行っていない. 本研究では, 機械学習の最大エントロピー法を利用して, 新聞データから日経平均の上昇下降を予測するとともに株式相場や経済に関わる知見を取得することで, 株式市場の予測に役立たせることを目指す.

本研究の主な主張点を以下に整理する.

- 機械学習を使用し，株価予測の実験を行った結果，分類先を 3 分類としたときの正解率は，提案手法が 0.451 であった. また, 日経平均 1 株あたりの 1 年間での平均利益は, 提案手法が 661 円, ベースラインが 1008 円であったが, 1 年毎の損失の最大値は提案手法が-1375 円, ベースラインが-3866 円であったため, この提案手法が有効な点も見られた.

- 実際に機械学習における素性 (学習に用いる情報のこと) を分析することで株式相場や経済に関わる重要な情報を把握することができ，それらに役立つ情報を明らかにした．例として株価上昇に関する素性には「ドル高」「金融緩和」などがあり，株価下降に関する素性には「利益確定」「円高」などがあった.

本論文の構成は以下の通りである．第 2 章では，本研究に関連する研究としてどのような研究が行われてきたかを記述し，その研究と本研究との関連を説明する．第 3 章では，本研究で扱う問題の設定とそれを解決するために提案した手法について説明を行う．第 4 章では，本研究で行った株価予測の実験についての説明と，その結果と考察について記述する．第 5 章では，本研究で行った素性分析の実験についての説明と, その結果と考察について記述する. 第 6 章では, 提案手法ではないその他の株価予測の実験についての説明と, その結果と考察について記述する. 第 7 章では, 今後の課題について記述する. 第 8 章ではまとめを行う.

第2章　先行研究

本章では，先行研究について記述する．2.1 節では，松井ら [1] が行った新聞記事の時系列テキスト分析による株式市場の動向予測について記述し、2.2 節では，中川ら [2] が行った出現頻度と連接頻度に基づく専門用語抽出について記述する．2.3 節では、赤江 [3] が行った機械学習を用いた類義語の使い分けに関する知識獲得について記述する．

2.1　新聞記事時系列テキストによる株式市場の動向予測

松井ら [1] は, 定期的に発行されるテキスト・データを時系列データと捉えることによって, テキストの差分に着目した分析を行い, 予測対象の動きを予測する．この研究ではこれを時系列テキスト分析と呼び, 新聞記事を対象とした時系列テキスト分析の手法を提案した．この手法は, 分析する時点のテキスト・データとその直前のテキスト・データを比較し, 新たに出現した語, 続けて出現している語, 消滅した語を特徴語として抽出して特徴ベクトルを作成し, SVM を用いてテキストの変化と市場の変化の関係を学習している．この結果を日本経済新聞の記事に適用し, 東証株価指数 (TOPIX) の日中の騰落を予測した．

評価実験では, 日本経済新聞を対象として, 予測対象日の前営業日の夕刊から予測対象日の朝刊までを一つのテキストとし, その見出しのみを用いた．予測対象は 2008 年から 2013 年までの東証株価指数 (TOPIX) 連動型上場投資信託 (ETF) とし, 予測対象日の寄りから引けにかけて TOPIX ETF の取引価格が上昇するか下落するか (終値が始値よりも高いか低いか) を予測した．訓練データの期間は, 予測対象日の直近の過去５年間とした．実験の結果, 全体の正解率は 0.714 であり, 最も予測精度が低い年でも正解率が 0.563 であった．また, この予測に基づく運用シミュレーションの平均年間利益率は 149 %であった．

2.2 出現頻度と連接頻度に基づく専門用語抽出

中川ら [2] はテキストデータから専門用語を取り出すため, 専門用語自動抽出システム (TermExtract) を作成した. TermExtract は名詞 (単名詞と複合名詞) を対象として専門用語抽出を行うシステムである. まず対象となるコーパスから専門用語の候補となる語を抽出し, 次に各候補語の専門用語としての重要度を計算する. その結果, スコアの高い順に候補語をソートしたものを出力している. なお, 重要度計算に単名詞バイグラムを用いることにより, 複合名詞がどのような単名詞で構成されているかという連接情報と候補語の頻度情報を専門用語かどうかの手掛かりとしている.

2.3 機械学習を用いた類義語の使い分けに関する知識獲得

赤江 [3] は, 教師あり機械学習を用いることにより, 類義語の使い分けを行い, 類義語の使い分けに関わる知見を得ている.

類義語 11 組について, 類義語の組ごとに類義語の使い分けの実験を行った. 入力文は, 1991 年〜1995 年, 2011 年〜2015 年の毎日新聞から獲得した, 類義語のいずれかの語を含む文である. 評価は 10 分割のクロスバリデーションで行う. 類義語の組のうち出現頻度が多かった語を全ての問題の分類先とするものをベースライン手法とし, 提案手法とベースライン手法の性能の比較を行う. 実験の結果, 正解率のマクロ平均は「データ数を出現率に合わせた実験」では, 提案手法が 0.84, ベースライン手法が 0.65 であり, 「データ数を同数に揃えた実験」では, 提案手法が 0.81, ベースライン手法が 0.42 であったため, この提案手法自体が類義語の使い分けに対して有用である.

この研究では, いくつかの類義語について実際に使い分けに役立ったと思われる情報を明らかにした. 特に, 類義語の使い分けに関する文献に載ってないような新たな知見が多く得られている. 例えば, 「作成」は「表」「リスト」などを作る時に使われ, 「作製」は「細胞」「遺伝子」などを作る時に使われるなどの素性を得られた. また, 品詞間における類義語の使い分けに関する特徴も得られた. この 2 つの成果は, 文章を生成する際の類義語の選択, 適切な表現の使い分けの提案に利用できる.

第3章　問題設定と提案手法

3.1　問題設定

先行研究 [1] では日経平均の上昇下降の予測のみを行っているが, 経済に関わる知見は取得していない. そこで, 本研究はそこに着目し, 機械学習を用いて予測とともに経済に関わる知見を取得していくことを目的とする.

3.2　株価予測の提案手法

本研究では, 教師あり機械学習を利用して、どのような語が記事中にあれば株価が上昇下降するのかを推定する. 株価の予測では当日の朝刊から当日の始値と終値の差の予測を行う. 入力には, 2007 年から 2018 年の毎日新聞のタイトル, 本文中の特定の単語 (前日終値比, 前日比など) を含む文を用いる. 株価が上昇, 変化なし, 下降をその文の分類先として機械学習を行う. また, 0.5 %以上の変動で上昇下降と定義する. 教師あり機械学習には最大エントロピー法を利用する.

3.3　知見獲得の提案手法

株式相場を予測する上で, どのような素性が役に立つのかを明らかにするために, 素性の分析を行う. 機械学習の分類性能が高い素性は特に役立つものである可能性がある. そこで, 機械学習の正解率を上げるために 2 日前の終値と前日の終値の差を推定する. 入力文には, 2007 年から 2018 年の毎日新聞のタイトル, 本文中の特定の単語 (前日終値比, 前日比など) を含む文を用い, 株価が上昇, 下降をその文の分類先として機械学習を行う. 正解率が最も高かった条件での素性を分析する. 最大エントロピー法では, 学習に役立つ素性が得られるため, 機械学習の素性を分析することで株式相場や経済にかかわる知見を獲得する.

3.4 最大エントロピー法

本研究では，教師あり機械学習法に，最大エントロピー法を使用する．

最大エントロピー法とは，あらかじめ設定しておいた素性 $f_i(1 \leq j \leq k)$ の集合を F とするとき，式 (3.1) を満足しながらエントロピーを意味する式 (3.2) を最大にするときの確率分布 $p(a,b)$ を求め，その確率分布にしたがって求まる各分類の確率のうち，もっとも大きい確率値を持つ分類を求める分類とする方法である [4, 5, 6, 7].

$$\sum_{a\in A,b\in B} p(a,b)g_j(a,b) = \sum_{a\in A,b\in B}(a,b)g_j(a,b) \quad for \quad \forall f_j(1 \leq j \leq k) \tag{3.1}$$

$$H(p) = -\sum_{a\in A,b\in B} p(a,b)log(p(a,b)) \tag{3.2}$$

ただし，A, B は分類と文脈の集合を意味し，$g_i(a,b)$ は 文脈 b に素性 f_i があってなおかつ分類が a の場合 1 となり，それ以外で 0 となる関数を意味する．また，(a,b) は，既知データでの (a,b) の出現の割合を意味する．

式 (3.1) は，確率 p と出力と素性の組の出現を意味する関数 g をかけることで 出力と素性の組の頻度の期待値を求めることになっており，右辺の既知データにおける期待値と，左辺の求める確率分布に基づいて計算される期待値が等しいことを制約として，エントロピー最大化 (確率分布の平滑化) を行なって，出力と文脈の確率分布を求めるものとなっている．

3.5 素性

本研究では, 素性は以下のものを用いる.

- TermExtract[2] により抽出した専門用語

素性は入力文とした文から取り出す.

第4章　株価予測の実験と考察

本章では, 本研究の株価予測の実験で使った実験データを 4.1 節で記述し, 本研究が行った実験方法を 4.2 節で記述し, 実験結果を 4.3 節に記述する. 4.4 節で実験の考察について記述する.

4.1　実験データ

本研究では, 実験データに 2007 年から 2018 年の毎日新聞のデータを用いる. テストデータは 1 年とし, 学習データはテストデータで用いた年の過去 3 年とする.

4.2　実験方法

3.2 節の方法を使って株価予測の実験を行う. 入力は, 以下の手法 1 から手法 3 とする.

手法 1　当日の朝刊のすべての記事のタイトル

手法 2　当日の朝刊の「前日比」「前日終値比」を含む段落

手法 3　当日の朝刊のすべての記事のタイトルと「前日比」「前日終値比」を含む段落

手法 3 では「前日比」「前日終値比」を含む段落が存在しない日は, 朝刊のタイトルのみで予測を行うこととする.

Buy&Hold(ここでは年始に買って年度末まで保持し続ける方法とする) をベースライン手法とし, 提案手法とベースライン手法の性能の比較を行う.

提案手法での利益の計算は, 株価が上昇と予測とした場合, 当日の始値で日経平均 1 株を買い, それを当日の終値で売る. 株価が下降と予測した場合, 当日の始値で日経平均 1 株を売り, をれを当日の終値で買い戻し, その差額を利益とする. また, 株価が変化なしと予測した場合は取引を行わないものとする.

4.3 実験結果と考察

手法 1 で行った実験の結果を表 4.1 に示す. 表 4.1 の各列は左から順に, 予測を行った年, 各年毎の正解率, データ総数, 利益 (total), 損益が最低となったときの値 (low), 買い注文の回数, 売り注文の回数, ベースライン手法の利益, ベースライン手法の損益が最低となったときの値である. 同様に手法 2 で行った実験の結果を表 4.2 に, 手法 3 で行った実験の結果を表 4.3 に示す. この結果をもとに有意差を検定した. 各年毎の利益と取引をしなかった場合 (損益 0) とを, 両側検定により有意差検定を行った. これらの結果を表 4.4 にまとめる. 手法 1 では p 値が 0.046 であり, 0.05 以下であるので有意差がみられた.

実験結果より, 新聞のタイトル情報だけでも予測はできることが分かった. 手法 2 では他の手法と比較して利益が少なくなっているが, これはデータ数が少なかったことが原因であると考えられる. データ総数が手法 1 と手法 3 では 2207 であるのに対し, 手法 2 では 819 であるため, これらの数を揃えた場合手法 2 の利益が手法 1 より多くなることも考えられる.

表 4.4 より, 提案手法はベースラインと比較して利益は劣っているが 1 年間の損失の最大値は小さくなっている. また, 手法 1 では有意差ありとなっているので, 提案手法でもある程度の予測はできているといえる.

表 4.1: 手法 1(すべての記事のタイトル)

年	正解率	総数 (U,D,N)	total	low	買い	売り	Buy&Hold	low(B&H)
10	0.346	243(57,63,123)	728	-409	14	96	-381	-1776
11	0.453	245(47,60,138)	597	-136	25	77	-1897	-2214
12	0.524	248(53,52,143)	39	-95	9	34	1835	-271
13	0.351	245(88,69,88)	-464	-464	7	2	5687	-119
14	0.414	244(51,60,133)	1722	-21	92	15	1303	-2178
15	0.443	244(63,52,128)	2825	-442	35	58	1708	-517
16	0.392	245(69,65,111)	-178	-1376	23	33	296	-3866
17	0.657	248(38,33,177)	15	-614	18	10	3466	-1059
18	0.477	245(55,54,136)	672	-702	34	18	-2959	-3711
平均	0.451		661				1006	

表 4.2: 手法 2(「前日比」「前日終値比」を含む段落)

年	正解率	総数 (U,D,N)	total	low	買い	売り	buy&hold	low(b&h)
10	0.393	89(22,24,43)	171	0	31	20	-381	-1776
11	0.300	100(20,26,54)	-197	-494	24	35	-1897	-2214
12	0.520	102(25,18,59)	-93	-275	18	13	1835	-271
13	0.365	107(44,27,36)	915	-431	28	18	5687	-119
14	0.361	86(19,20,47)	240	0	26	10	1303	-2178
15	0.396	91(24,16,51)	702	-451	39	3	1708	-517
16	0.462	104(30,30,44)	1844	-402	25	12	296	-3866
17	0.712	66(7,8,51)	710	0	3	7	3466	-1059
18	0.432	74(12,23,39)	-1775	-1775	15	19	-2959	3771
平均	0.438		279				1006	

表 4.3: 手法 3(すべての記事のタイトルと「前日比」「前日終値比」を含む段落)

年	正解率	総数 (U,D,N)	total	low	買い	売り	buy&hold	low(b&h)
10	0.362	243(57,63,123)	1034	-409	18	90	-381	-1776
11	0.457	245(47,60,138)	892	-87	21	78	-1897	-2214
12	0.520	248(53,52,143)	36	-114	8	32	1835	-271
13	0.351	245(88,69,88)	-179	-230	8	3	5687	-119
14	0.414	244(51,60,133)	2016	-96	91	11	1303	-2178
15	0.447	244(63,52,128)	3355	-49	43	55	1708	-517
16	0.388	245(69,65,111)	-871	-1716	25	33	296	-3866
17	0.661	248(38,33,177)	-357	-734	16	11	3466	-1059
18	0.478	245(55,54,136)	1077	-432	32	18	-2959	-3711
平均	0.453		778				1006	

表 4.4: 各手法での実験結果

入力文	正解率	総数	1 年間の 平均利益 (円)	1 年間の損失 最大値 (円)	t 検定 p 値
手法 1	0.451	2207	661	-1375	0.046
手法 2	0.438	819	279	-1775	0.207
手法 3	0.453	2207	778	-1716	0.057
ベースライン			1006	-3866	0.143

第5章　知見獲得の実験と考察

本章では, 本研究の素性分析の実験方法を 5.1 節で記述し, 実験結果を 5.2 節に記述する. 5.3 節で実験の考察について記述する.

5.1　実験方法

本研究の評価は 10 分割クロスバリデーションで行う. 3.3 節の方法を使って当日の朝刊から 2 日前の終値と前日の終値の差の推定を行い, 素性分析の実験を行う. 入力文は, 以下の手法 1 から手法 4 とする.

手法 1 当日の朝刊のすべての記事のタイトル

手法 2 当日の朝刊の「前日比」「前日終値比」を含む段落

手法 3 当日の朝刊の「前日比」「前日終値比」を含み, かつ「日経平均」を含む段落

手法 4 当日の朝刊の「日経平均」を含む段落

5.2　実験結果と考察

5.1 節の手法 1 から手法 4 で行った実験の正解率を表 5.1 に示す. 分類先は上昇・下降の 2 分類であり, 正解率は 10 分割のクロスバリデーションでもとめた.

表 5.1: 各手法での正解率

入力文	正解率
手法 1	0.515
手法 2	0.711
手法 3	0.953
手法 4	0.889

表 5.1 からわかるように, 2 日前の終値と前日の終値の株価の差を朝刊のタイトルから推定した場合, 正解率が 0.515 と低くなった. 今回の手法では, 分類先を 2 分類としたため正解率が 0.5 近辺となった場合は推定ができていないといえる. これより, 新聞記事のタイトルは前日の株価の変動とは関連性が少ないことが考えられる. 対して「前日比」「前日終値比」を含む段落は前日の株価の変動が記事となったものが多いため, 正解率は高くなった. 同時に「日経平均」を含む段落を入力文としたときはさらに正解率が上がった. これは, 「前日比」「前日終値比」を含む記事では日経平均株価以外の騰落を記事にしたものがあるため, 「日経平均」という単語を追加したことにより, それらの記事が入力文から外れたためだと考えられる.

5.3 素性の分析

正解率が最も高かった 5.1 節の手法 3 で用いた学習データから素性の分析を行う.

5.3.1 正規化 α 値

正規化 α 値とは, 最大エントロピー法で求まる α 値を全分類先での合計が 1 となるように正規化した値である. また, 素性 a と分類先 b の対によって定まる値であり, 素性 a のみが適用される場合に分類先 b となる確率に相当する. 各素性の, 分類先ごとに与えられた正規化 α 値が高いほど, その分類先であることを推定するのに重要な素性であることを意味する.

5.3.2 有用素性

機械学習に使った素性のうち株価上昇の正規化 α 値の上位 30 個を表 5.2 に, 下位 30 個を表 5.3 に示す. 正規化 α 値が上位のものは株価上昇に役立つことが, 下位のものは株価下降に役立つことがわかる.

株価上昇に関する有用な素性を人手で考察し, その上昇, 下降の頻度を調べた. その一部を表 5.4 に示す. 同様に株価下降に関するものを表 5.5 に示す.

日経平均株価は下降することよりも上昇する頻度の方が多かったため, 上昇下降のどちらともとれる素性は上昇の正規化 α 値が高くなる傾向にあった. よって株価上昇に関する知見を獲得することは, 株価下降に関する知見を獲得することよりも困難であった.

表 5.2: 正規化 α 値の上位 30 個)

	α 値	
高	0.016580	0.983420
反発	0.135390	0.864610
続伸	0.167425	0.832575
発	0.195921	0.804079
高値	0.196211	0.803789
買い	0.207213	0.792787
ドル高	0.242436	0.757564
感	0.274019	0.725981
東京株	0.293546	0.706454
連続	0.319628	0.680372
上げ幅	0.325236	0.674764
米	0.326754	0.673246
鈴木一也	0.335755	0.664245
銘柄	0.336227	0.663773
回復	0.344377	0.655623
局面	0.346091	0.653909
上海総合指数	0.354018	0.645982
発表	0.355278	0.644722
時点	0.355572	0.644428
業績	0.363781	0.636219
自動車	0.373515	0.626485
円	0.374044	0.625956
ＥＵ	0.374429	0.625571
各国	0.379571	0.620429
ー	0.380039	0.619961
機	0.382057	0.617943
格好	0.382489	0.617511
首脳	0.383072	0.616928
伸し	0.384034	0.615966
失望	0.386034	0.613966

表 5.3: 正規化 α 値の下位 30 個)

	α 値	
反落	0.922276	0.077724
安	0.897168	0.102832
続落	0.834570	0.165430
下落	0.792641	0.207359
連休	0.782734	0.217266
利益確定	0.779944	0.220056
日本市場	0.758697	0.241303
リスク	0.710041	0.289959
株式	0.709117	0.290883
価格	0.704095	0.295905
利回り	0.700072	0.299928
下げ	0.699317	0.300683
悪化懸念	0.694002	0.305998
長期金利	0.693954	0.306046
利益	0.693584	0.306416
円高	0.691278	0.308722
日銀	0.687081	0.312919
場面	0.682163	0.317837
年	0.680008	0.319992
超	0.676204	0.323796
債券	0.674759	0.325241
懸念	0.670320	0.329680
本格化	0.666414	0.333586
日本	0.666325	0.333675
殺到	0.657257	0.342743
トランプ	0.653167	0.346833
下げ幅	0.652636	0.347364
田所柳子	0.649101	0.350899
東証株価指数	0.645429	0.354571
警戒感	0.632569	0.367431

表 5.4: 有用素性の考察 (上昇)

素性	上昇	下降	考察
ドル高	46	7	円安・ドル高により株価上昇
回復	119	15	株価が○○○円台を回復した 景気回復への期待感から株価上昇
業績	32	13	業績改善への期待、企業業績改善への期待、好業績などから株価上昇
自動車	13	7	自動車などの輸出関連株を中心に買われ、売られ 円安の時に買われ、円高のときに売られる 上昇が多いのはたまたま
営業日	122	82	○営業日連続で上昇、下落 ○営業日ぶりに上昇、反発、反落 上昇が多いのは日経平均株価が上がることの方が多いから
為替相場	12	7	為替相場の円安により上昇 為替相場が円高に振れたことで下降
米株高	26	0	前日の米株高を好感して株価上昇
金融緩和	24	16	金融緩和により株価上昇

表 5.5: 有用素性の考察 (下降)

素性	上昇	下降	考察
利益確定	4	18	利益確定の売りに押され、売り注文が優勢となり等、上昇した後の反動で株価下降
リスク	4	23	リスクを回避する姿勢が強まり等、リスクを回避するために売りが強まり株価下降
株式	316	267	東京株式市場がほとんど 下降が多いのはたまたま
価格	2	13	原油価格の下落により株価が下降すること多かった
利回り	2	11	国債の利回りの増減 利回りの増減と株価の相関は見られなかった
悪化懸念	1	13	企業業績の悪化懸念が強まり株価下降
長期金利	6	9	長期金利の増減 長期金利の指標となる 10 年債の利回りの増減 長期金利が下がれば株価も下降
利益	6	9	利益を確定する売りにより株価下降
円高	36	109	円高により株価下降
不透明感	3	20	経済の先行き不透明感などで株価下降

第6章　条件を変えた場合の実験と考察

本研究では, 株価の予測を行ったがその提案手法については 4.2 節に記述した通りである. この手法に至る過程で入力文や学習データの年数を変えるなど, 他にも様々なパターンで行った. 本章では, それらの実験の一部を記述する.

6.1　実験方法

3.2 節の方法を使って株価予測の実験を行う. 入力文は以下の手法 1, 手法 2 として, 学習データを 1 年, 5 年として実験を行った.

手法 1　当日の朝刊のすべての記事のタイトル

手法 2　当日の朝刊の「前日比」「前日終値比」を含む段落

6.2　実験結果

手法 1, 学習データ 1 年で行った実験の結果を表 6.1 に示す. 表 6.1 の各列は左から順に, 予測を行った年, 各年毎の正解率, データ総数, 利益 (total), 損益が最低となったときの値 (low), 買い注文の回数, 売り注文の回数, ベースライン手法の利益, ベースライン手法の損益が最低となったときの値である. 同様に手法 1, 学習データ 5 年で行った実験の結果を表 6.2 に, 手法 2, 学習データ 1 年で行った実験の結果を表 6.3 に, 手法 2, 学習データ 5 年で行った実験の結果を 6.4 示す.

表 6.1: 手法 1(すべての記事のタイトル) 学習データ 1 年

年	正解率	総数 (U,D,N)	total	low	買い	売り	buy&hold	low(b&h)
08	0.339	245(90,89,66)	3193	-194	28	128	-6296	-7529
09	0.354	242(73,82,87)	830	-60	146	81	1555	-1932
10	0.387	243(57,63,123)	580	-154	54	19	-381	-1776
11	0.539	245(47,60,138)	157	-102	15	13	-1897	-2214
12	0.577	248(53,52,143)	0	0	0	0	1835	-271
13	0.363	245(88,69,88)	595	-153	9	8	5687	-119
14	0.336	244(51,60,133)	1237	-277	106	11	1303	-2178
15	0.496	244(63,52,128)	-151	-162	0	13	1708	-517
16	0.429	245(69,65,111)	-337	-727	14	2	296	-3866
17	0.589	248(38,33,177)	115	-460	48	12	3466	-1059
18	0.555	245(55,54,136)	0	0	0	0	-2959	-3711
平均	0.456		565				392	

手法 1 で行った実験の結果を表 6.5 に, 手法 2 で行った実験の結果を表 6.6 にまとめる. 学習データ 3 年での実験は 4 章で行ったものである.

表 6.1 より, 手法 1, 学習データ 1 年で行った実験は利益が出ている年が多かったが取引なしで利益が 0 となっている年が見られた. これは学習データが少なかったために正しく予測が行えていなかったことが原因と考えられる. また, 表 6.5 より, 手法 1 では学習データを 5 年としたときは, 他のものよりも正解率, 利益ともに低くなっている. よって学習データを増やしすぎても正解率が上がることはない. これは各年ごとで流行や傾向が違うため, 古すぎる情報は株価の予測に役立たないためと考えられる. 手法 2 では, 学習データを増やすほど利益は多くなっていたが, これは手法 1 と比較してデータ総数が少なく, 学習データの量が不足していたためと考えられる. また, 学習データを 5 年以上に増やした場合, これ以上利益が増えることはなかった. 以上より, 学習データは 3 年近辺が一番予測の性能が高くなることがわかった.

表 6.2: 手法 1(すべての記事のタイトル) 学習データ 5 年

年	正解率	総数 (U,D,N)	total	low	買い	売り	buy&hold	low(b&h)
12	0.444	248(53,52,143)	-395	-600	32	54	1835	-271
13	0.343	245(88,69,88)	-2243	-2293	36	12	5687	-119
14	0.373	244(51,60,133)	830	-21	110	11	1303	-2178
15	0.414	244(63,52,128)	2565	-218	58	71	1708	-517
16	0.420	245(69,65,111)	-413	-466	11	13	296	-3866
17	0.581	248(38,33,177)	-409	-469	13	26	3466	-1059
18	0.453	245(55,54,136)	-29	-1168	42	46	-2959	-3711
平均	0.431		-13				1619	

表 6.3: 手法 2(「前日比」「前日終値比」を含む段落) 学習データ 1 年

年	正解率	総数 (U,D,N)	total	low	買い	売り	buy&hold	low(b&h)
08	0.3143	245(90,89,66)	643	-694	43	35	-6296	-7529
09	0.2899	242(73,82,87)	16	-602	25	20	1555	-1932
10	0.3820	243(57,63,123	478	-56	22	27	-381	-1776
11	0.3000	245(47,60,138)	178	-216	20	26	-1897	-2214
12	0.5686	248(53,52,143)	544	0	25	18	1835	-271
13	0.3364	245(88,69,88)	-1615	-1627	44	27	5687	-119
14	0.2907	244(51,60,133)	-541	-904	19	20	1303	-2178
15	0.4945	244(63,52,128)	1794	0	24	16	1708	-517
16	0.3942	245(69,65,111)	-400	-996	30	30	296	-3866
17	0.3485	248(38,33,177)	1	-307	7	8	3466	-1059
18	0.5135	245(55,54,136)	-182	-341	12	23	-2959	-3711
平均	0.3848	83	-480				392	

表 6.4: 手法 2(「前日比」「前日終値比」を含む段落) 学習データ 5 年

年	正解率	総数 (U,D,N)	total	low	買い	売り	buy&hold	low(b&h)
12	0.441	102(25,18,59)	145	-344	32	18	1835	-271
13	0.421	107(44,27,36)	3644	-83	38	21	5687	-119
14	0.314	86(19,20,47)	845	-499	40	18	1303	-2178
15	0.385	91(24,16,51)	1033	-805	29	12	1708	-517
16	0.442	104(30,30,44)	619	-861	28	8	296	-3866
17	0.636	66(7,8,51)	147	-78	5	8	3466	-1059
18	0.460	74(12,23,39)	-2461	-2461	19	13	-2959	3771
平均	0.443		567				1619	

表 6.5: 手法 1 の学習データ別の実験結果

学習データ (年)	正解率	1 年間の 平均利益 (円)	1 年間の損失 最大値 (円)
1	0.456	565	-727
3	0.451	661	-1376
5	0.431	-13	-2293

表 6.6: 手法 2 の学習データ別の実験結果

学習データ (年)	正解率	1 年間の 平均利益 (円)	1 年間の損失 最大値 (円)
1	0.385	83	-1627
3	0.451	279	-1775
5	0.431	567	-2461

第7章　今後の課題

本研究では, 教師あり機械学習を用いて株価の予測と, 株式相場や経済に関わる知見の獲得を行ったが, いくつかの問題が残っている. 本章では, その問題を今後の課題として以下にまとめる.

- 本研究で行った株価の予測であるが, 提案手法で一番利益が高かったものの正解率は 0.451 であった. これはベースラインと比較して損失の最大値は抑えらているものの, 利益は劣っている. 入力や素性を他のものに変え, 予測の正解率が向上するものを探る.
- 知見獲得の素性分析に使用した学習データは, 2 日前の終値と前日の終値の騰落を当日の朝刊から予測したものであり, 実際に株価予測に使用した学習データではなかった. よって得られた知見は株価の騰落に関するものなど, 単純なものが多かった. 正解率を向上させた株価予測の学習データを用いることでより興味深い知見を獲得することを目指したい.
- 本研究では, 毎日新聞から株価の予測を行ったが, 他の新聞記事やウェブデータなど, 他のデータから株価が予測できるのかを調査したい.

第8章　おわりに

本研究では, 教師あり機械学習を用いて株価予測と, 株式相場や経済に関わる知見獲得を行った.

株価の予測では, 毎日新聞を学習データとテストデータに用いて実験を行った. 実験の結果, 新聞記事のタイトルを用いるだけでもある程度の予測ができることがわかった. 提案手法では, 利益はベースラインより劣ったが, 1 年間の損失の最大値はベースラインよりも小さかった. よって, 使い方によっては提案手法が役立つ場合もあると考えられる.

知見獲得の研究では, 機械学習に使った学習データの素性を分析することで, 株価の騰落に関わる知見を獲得することができた. しかし, 本研究の実験において知見獲得のために用いた学習データは実際に株価予測に使ったものではなかった. よって, 得られた知見は単純なものが多かった. より興味深い知見を得るためには, 素性分析を行う学習データを実際に株価予測に使ったものにする必要があると考えられるが, そのためには予測の正解率を向上させる必要がある.

謝辞

本研究を進めるに当たり, 研究の進め方や本論文の書き方など, 細部にわたる御指導を頂きました, 鳥取大学工学部電気情報系学科自然言語処理研究室の村田真樹教授に心から御礼申し上げます. また, 本研究を進めるにあたり, 御指導, 御助言を頂きました, 村上仁一准教授に心から御礼申し上げます. その他様々な場面で御助言を頂いた自然言語処理研究室の皆様に感謝の意を表します.

参考文献

[1] 松井藤五郎, 和泉潔. 新聞記事の時系列テキスト分析による株式市場の動向予測. 人工知能学会第 30 回全国大会論文集, 2016.

[2] 中川裕志, 森辰則, 湯本紘彰. 出現頻度と連接頻度に基づく専門用語抽出. 自然言語処理, Vol. 10, No. 1, pp. 27–45, 2003.

[3] 赤江涼太. 機械学習を用いた類義語の使い分けに関する知識獲得. 卒業論文, 鳥取大学工学部知能情報工学科, 2018.

[4] Eric Sven Ristad. Maximum entropy modeling for natural language. In *ACL/EACL Tutorial Program, Madrid*, 1997.

[5] 村田真樹, 内山将夫, 内元清貴, 馬青, 井佐原均. 種々の機械学習手法を用いた多義解消実験. 電子情報通信学会言語理解とコミュニケーション研究会, NLC2001-2, pp. 7–14, 2001.

[6] Masao Utiyama. Maximum entropy modeling packagen: http://www.nict.go.jp/x/x161/members/mutiyama/software.htmlmaxent, 2006.

[7] Masaki Murata, Kiyotaka Uchimoto, Masao Utiyama, Qing Ma, Ryo Nishimura, Yasuhiko Watanabe, Kouichi Doi, and Kentaro Torisawa. Using the maximum entropy method for natural language processing: Category estimation, feature extraction, and error correction. *Cognitive Computation*, Vol. 2, No. 4, pp. 272–279, 2010.