ニュース重要単語の機械学習によるアクティブ運用

Active Investment Management by Machine Learning of Important News Keywords

三好勝博[1*]　細木唯以[1]　江口潤一[2]　鈴木智也[1,2]
Masahiro Miyoshi[1] Yui Hosoki[1] Junichi Eguchi[2] Tomoya Suzuki[1,2]

[1] 茨城大学大学院理工学研究科
[1] Graduate School of Ibaraki University
[2] 大和証券投資信託委託株式会社
[2] Daiwa Asset Management Co.Ltd.

Abstract: In stock investment and asset management, numerical information has been mainly used for fundamentals analysis and technical analysis, but recently text information such as news released in real time can be also used by development of natural language processing techniques. Although machine learning is mainly used for investment decisions, because text information is composed of a large number of words, its bag-of-words tend to be a sparse and high-dimensional vector. Therefore, the curse of dimensionality causes a negative effect on the learning performance of machine learning. For this reason, we only focus on news headlines to restrict the number of keywords as text information. In addition, we extracted important keywords that increase the volatility of stock prices right after the news appears, and applied machine learning techniques to learn the relationship between the combination of important keywords included in a news headline and tomorrow's active return of the company most related to the news. Through some statistical tests, we could confirm the validity of focusing on the stock volatility to extract important keywords and their combination is useful for active investment management with machine learning approach.

1　はじめに

近年では AI による自然言語処理技術が発展し，イベント的に発生するニュースなどのテキスト情報もアルゴリズムに組み込んだ投資判断が活発になりつつある [1, 2, 3]．投資判断には主に機械学習を用いるが，テキスト情報は多数のワードで構成されているため Bag-of-words の局所表現によりベクトル化すると高次元でスパースになりやすい．その結果，次元の呪いにより機械学習の学習性能に悪影響を及ぼす．

そこで本研究では，キーワードの種類を抑制するため，テキスト情報の中から将来の株価変動 (ボラティリティ) に大きく影響する重要キーワードを抽出する．その後，抽出した重要キーワードを用いて，各ニュースヘッドラインに含まれる重要キーワードの組み合わせと翌日のアクティブリターンを機械学習し，アクティブ運用による投資シミュレーションを行う．なおアクティブ運用とは，ベンチマーク (本研究では TOPIX) を上回る運用パフォーマンスを目指す運用手法である．

*連絡先：茨城大学理工学科機械システム工学専攻
〒 316-8511 茨城県日立市中成沢町 4-12-1
E-mail: 18nm500x@vc.ibaraki.ac.jp

2　ニューステキストの前処理

本研究では使用するテキスト情報として，QUICK ニュース解析サービス [4] から配信されるニュースヘッドラインを用いる．形態素解析ツールである MeCab[5] を用いてニュースヘッドラインを形態素に分割し，重要キーワードの候補を得る．ヘッドラインはニュース本文と比べテキスト量が少ないが，重要な情報およびキーワードが凝縮されている．さらにテキスト情報をヘッドラインに限定することで，登場するキーワードを抑制する．

ヘッドライン中には●や★などの装飾記号や数字が混在し，ワード数の増加を招くと考えられる．そのため，予めこれらのワードを削除する．また本研究では，個別銘柄に対してアクティブ運用を行うため，ニュースの配信元が「TECH」「QEC」「日銀」「財務省」であるヘッドラインや，銘柄コードに「日銀」「JPX」「日証金」を含むヘッドラインは対象外とする．

ヘッドライン中には共起関係の高いキーワード (共起ワード) が存在する．共起関係とは，ある単語がある文章中に出現した際に，その文章中に特定の単語が頻出することを指す．例えばエアリズムなど，特定の銘柄 c (この場合はユニクロ) のみに対応するキーワードは，株価変動との関係を汎用的に評価できない．そこで，式 (1) を満たすキーワード k を共起ワードとみなし，予めこれらも削除する．

$$P(c|k) = \frac{N(c|k)}{N(k)} \geq 0.5 \tag{1}$$

ここで $N(k)$ はキーワード k を含むニュース数，$N(c|k)$ はキーワード k を含むニュースが銘柄 c と最も関連した回数である．つまり式 (1) は，キーワード k を含むニュースの半分以上が銘柄 c のものであることを意味する．なお，銘柄コードが付与されていないニュースは c が不明なので対象外とする．

3　重要キーワードの抽出

3.1　ボラティリティ変化率

本研究では，株価ボラティリティに注目し，重要キーワードを抽出する．金融市場に大きな影響を与えるニュースが出現すると，そのニュースに関連する株価は大きく変動することが予想される．そこで本研究では，ニュース出現前後の株価ボラティリティの変化を計算し，ニュースヘッドラインを構成するキーワードの重要性を評価する．このボラティリティ変化が高いほど，重要なキーワードとみなす．

まず，出現したニュースと関連がある企業の株価を $x(d)$ とする．本研究では日次終値を用い，d は日付を意味する．株価の収益率 $r(d)$ は

$$r(d) = \frac{x(d) - x(d-1)}{x(d-1)} \tag{2}$$

であり，直近 D 日間の標準偏差を計算することで，ボラティリティ $v(d)$ を得る．

$$v(d) = \sqrt{\frac{1}{D}\sum_{i=0}^{D-1}\{r(d-i) - m(d)\}^2} \tag{3}$$

ここで $m(d)$ は直近 D 日間の収益率の平均値である．なお本研究では，1 カ月の営業日として $D = 20$ と設定した．

ニュース発生前後のボラティリティ変化を評価する際に，他媒体で既にニュースが報道された可能性を考慮するため，時間を遡るパラメータ b を導入する．ニュース発生前のボラティリティを $v(d-b)$ とすると，ニュース発生前後のボラティリティ変化 $V(d)$ は，

$$V(d) = \frac{v(d)}{v(d-b)} \tag{4}$$

として評価される．発生したニュースを構成する全キーワードに対して，この変化率 $V(d)$ を紐づける．

実際のニュースヘッドラインはリアルタイムに配信されるが，ニュースが配信されるたびに投資判断および売買注文を自動化するには，高額な設備投資が必要である．そこで本研究では試作的な運用として，毎日 12 時に直近 24 時間分のニュースデータを一括で取得し，得られたニュースに基づいて当日終値付近で売買を執行する．したがって，当日 d 日で得られるニュースには，前日 $d-1$ 日のニュースの一部が含まれている．よって，式 (4) に示すボラティリティ変化率 $V(d)$ の計算において $b \geq 2$ にする必要がある．他媒体での先行発表も考慮すると b を延長する必要があるが，本研究では最も基礎的な設定として $b = 2$ とする．

3.2　重要キーワードの検定

時間経過とともに様々ニュースが発生するため，各キーワード k $(k = 1, 2, \cdots, K)$ において変化率 V に関する集合 $\mathbf{G}_k$ を得る．この真の母集団は未知であるが，母平均を μ_k，母分散を σ_k^2 と書く．

まず，集合 $\mathbf{G}_k$ の平均値 $\bar{V}_k$ は母平均に関する不偏推定量となるため，$\bar{V}_k$ をキーワード k におけるボラティリティ変化率と考えることができる．しかし各キーワードにおいて出現回数 n_k は異なり，n_k が小さいほど推定値の信頼性は低下する．つまり $\bar{V}_k$ の大小を単純に比較することで，重要キーワードを選出できるとは言い難い．

そこで，母平均の検定を利用する．中心極限定理より，平均値 $\bar{V}_k$ は $N\left(\mu_k, \sqrt{\frac{\sigma_k^2}{n_k}}\right)$ に従う．キーワードとして重要性が無く，ボラティリティ変化が無い $(\mu_k = 1)$ ことを帰無仮説 H_0 とすると，帰無仮説 H_0 と集合 $\mathbf{G}_k$ のマハラノビス距離によって，検定統計量である t 値を得る．

$$t_k = \frac{\mu_k - 1}{\sqrt{\frac{\sigma_k^2}{n_k}}} \tag{5}$$

ただし μ_k と σ_k^2 は未知であるため，それぞれ平均値 $\bar{V}_k$ および不偏分散 $s_k^2 = \frac{1}{n_k - 1}\sum_{j=1}^{n_k}(V_{k,j} - \bar{V}_k)^2$ を用いる．この t_k に基づいて大小比較すれば，サンプルサイズ n_k の違いを考慮できる．本研究では有意水準 5% (片側検定) を満たすキーワード k に厳選する．

ただし n_k が少なすぎるキーワードについては，先に除去しておくことが賢明である．そこで本研究では，

10 日間に 1 回以上出現することを採用の目安とした．例えば，4 年間のニュースを解析対象にする場合，約 1000 営業日において出現回数が 100 回以下のキーワードは採用しない．1 年間のニュースを解析対象にする場合は，約 300 営業日において出現回数が 30 回以下のキーワードは採用しない．

3.3 実験

本研究では 2011 年から 2016 年に QUICK ニュース解析サービスから配信されたニュースヘッドラインをテキスト情報として使用した．キーワードの重要性は，流行や経済状況によって時間変化すると考えられるため，重要キーワードを抽出するニュースの参照期間を変えながら，株価に対するボラティリティを高めるキーワードを抽出し，キーワードリストを作成する．本研究で用いるニュースの参照期間として，過去 4 年間または 1 年間のように変更し，抽出される重要キーワードの違いを調査する．

表 1 は過去 4 年間及び過去 1 年間のニュースを参照した場合における重要キーワード上位 10 ワードを示す．結果として，参照期間 4 年間においては「円」や「東証」,「続伸」など金融市場特有のキーワードが多く見られる．一方，参照期間 1 年間においても金融市場特有のキーワードが多く見られるが,「マザーズ」や「東北」などの固有名詞やイベントワードが重要視される傾向にある．つまりニュース参照期間を長期にすることで，金融市場特有の一般的なキーワードが，短期にするとトレンドワードが抽出されやすい．今回は過去 4 年間と過去 1 年間のニュースヘッドラインを参照したが，より期間を短期に (3 カ月や 1 カ月) することで，より顕著にトレンドワードを抽出できると考えられる．また表 1 および表 2 において,「続伸」や「減」のように客観的にもボラティリティに影響すると考えられるキーワードが自動抽出されている．

表 1: 株価ボラティリティに影響を与える上位 10 ワード．参照期間 4 年間 (左列)，参照期間 1 年間 (右列)

2011~2014 年 (選出数 459)		2014 年 (選出数 282)	
キーワード k	t_k	キーワード k	t_k
中国	63.74	東証	50.96
続伸	60.95	輸出	49.73
反発	55.79	値がさ株	47.02
純	55.78	マザーズ	41.51
輸出	55.48	派遣	39.23
減	54.66	円	39.24
大証	54.59	先行	36.70
安心	52.35	格上げ	36.54
円	50.70	米	35.49
超	49.66	社長	34.92

2012~2015 年 (選出数 500)		2015 年 (選出数 268)	
キーワード k	t_k	キーワード k	t_k
円	96.29	東証	52.13
米	71.40	原油	46.74
中国	69.81	三菱	46.28
減	61.01	脱	43.80
大幅	59.82	発売	42.96
大証	59.33	中国	38.58
売る	58.78	純	37.44
原油	56.77	新潟	37.35
純	56.73	インフラ	37.34
反発	55.82	軟調	36.03

2013~2016 年 (選出数 500)		2016 年 (選出数 306)	
キーワード k	t_k	キーワード k	t_k
東証	112.99	東証	70.83
円	101.71	円	55.67
純	64.11	合併	53.41
減	62.42	純	52.46
米	61.78	公表	50.21
原油	61.38	買い戻す	48.30
安	59.27	海運	45.23
中国	57.73	一段	44.61
大幅	56.05	響く	42.05
反発	54.95	安	40.42

4 アクティブリターンの機械学習

4.1 学習方法

前章で作成したキーワードリストや，新規にリアルタイムに得られるニュースヘッドラインや株式価格情報を用いて，f 日後のアクティブリターン (銘柄 c と TOPIX との差) を予測する．現在を d とすると，f 日後のアクティブリターンは次式となる．

$$r_c^*(d+f) = \frac{x_c(d+f) - x_c(d)}{x_c(d)} - \frac{x_t(d+f) - x_t(d)}{x_t(d)} \tag{6}$$

ここで，x_c は銘柄 c の株価，x_t は TOPIX とする．

図 1 に各期間の役割を示す．緑色は 3 章にて抽出したキーワードリストを作成する期間であり，水色はフォワードテストを行うシミュレーション期間である．黒い太枠はキーワードリストを 1 年ごとに更新し，同期間のシミュレーションを行う．本研究では①～④の 4 パターンの期間においてシミュレーションを行う．

機械学習の手順は次の通りである．まず，ニュース

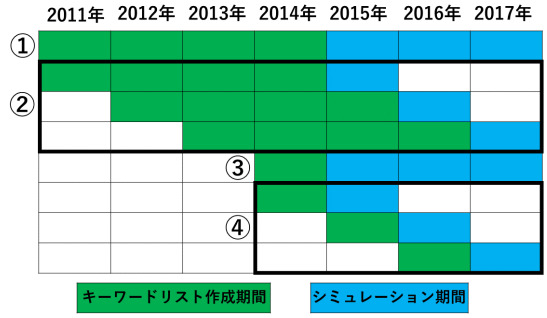

図 1: 各期間の役割 (全 4 パターン)

出現直後のアクティブリターン r_c^* に応じて，各ニュースに教師ラベルを付与する．なお各ニュースに最も関連する銘柄を c とする．

- if $r_c^*(d+f) > 0$, then ニュースの教師ラベル $= 1$
- if $r_c^*(d+f) < 0$, then ニュースの教師ラベル $= -1$
- if $r_c^*(d+f) = 0$, then ニュースの教師ラベル $= 0$

なお $r_c^*(d+f) = 0$ のケースは極めて少ないため，$r_c^*(d+f)$ に関する正・負の 2 クラス判別問題とみなす．

次に，3 章で抽出された重要キーワードを構成要素とする Bag-of-words によって，ニュースヘッドラインをベクトル $\mathbf{v}_c(d)$ として表現する．例えば，d 日にて銘柄 c に関連するニュースヘッドラインが「新作」「株価格付」「黒字」といった重要キーワードを含む時，これらのキーワードに対応するベクトル要素を 1 とし，他を 0 とする．このベクトル $\mathbf{v}_c(d)$ と，将来の $r_c^*(d+f)$ に基づく教師ラベルとの関係性を Support Vector Machine (SVM)[6] によって機械学習する．このように重要キーワードに限定することで，Bag-of-words の高次元化を抑制している．

図 1 のシミュレーション期間において，次のように学習と予測を繰り返す．まず，過去 s 日間 $(d-f-s \sim d-f)$ に配信されたニュースヘッドラインとその教師ラベルを用いて SVM を学習する．次に，当日 d に配信されたニュースヘッドラインを学習済み SVM に入力することで，f 日後のアクティブリターンを予測する．日付 d を進めながら毎日予測を行うが，SVM の再学習は s 日毎とする．なお本研究では，$s = 60$ および $f = 1$ とする．

シミュレーション期間における予測精度を表 2 と表 3 に示す．2 クラス判別問題であるため 50%が正答率の基準となるが，いずれの期間でも基準を超える正答率を示している．しかし $r_c^*(d+f) < 0$ と予測した場合の再現率は高いものの，$r_c^*(d+f) > 0$ と予測した場合の再現率は非常に低い．おそらく相対的にネガティブに作用するニュースが多いため，予測結果が負に偏ったと考えられる．しかし表 3 のように，重要キーワード

表 2: 重要キーワードを 1 つ以上含むニュースヘッドラインに限定した場合

		$r_c^*(d+f) > 0$ と予測			$r_c^*(d+f) < 0$ と予測		
期間	正答率	再現率	適合率	F 値	再現率	適合率	F 値
①	52.8%	1.0%	53.3%	2.0%	99.2%	52.8%	68.9%
②	52.7%	4.9%	48.9%	8.9%	95.4%	52.8%	68.0%
③	52.8%	1.8%	49.6%	3.5%	98.3%	52.9%	68.8%
④	52.6%	7.0%	48.9%	12.2%	93.4%	52.9%	67.5%

表 3: 重要キーワードを 3 つ以上含むニュースヘッドラインに限定した場合

		$r_c^*(d+f) > 0$ と予測			$r_c^*(d+f) < 0$ と予測		
期間	正答率	再現率	適合率	F 値	再現率	適合率	F 値
①	53.1%	5.3%	49.5%	9.6%	95.3%	53.4%	68.4%
②	52.5%	13.4%	48.7%	21.1%	87.4%	53.1%	66.0%
③	52.9%	5.5%	48.4%	9.9%	94.8%	53.2%	68.2%
④	52.5%	14.1%	49.6%	22.0%	87.3%	53.3%	66.2%

数を増やすほど再現率の偏りは改善されている．つまり SVM の学習において，重要キーワードの組み合わせによってニュース全体のボジネガを判断することが重要であり，各キーワードには非線形的な関係があると言える．

5 重要キーワードの妥当性

前章の機械学習においてキーワードの組み合わせが重要であるという知見を得たが，そもそもボラティリティ変化に基づいて重要キーワードを選出する妥当性は不明である．そこで 3 章で抽出された重要キーワードリストを「オリジナルリスト」とし，オリジナルリストに選抜されなかったキーワードをランダムに抽出することで「ランダムリスト」を 100 セット作成する．なおランダムリストに含まれるキーワード数は全てオリジナルリストと同数にする．次にオリジナルリストとランダムリストのそれぞれで Bag-of-words を構成し，各シミュレーション期間において SVM の学習および予測を行った．

それぞれの正答率を図 2〜図 5 に示す．なお，重要キーワードを 3 ワード以上含むニュースヘッドラインに限定した場合である．いずれの期間においても，オリジナルリストによる正答率はランダムリストによる正答率の中央値を超えており，重要キーワードの抽出においてボラティリティ変化に着眼する妥当性を確認

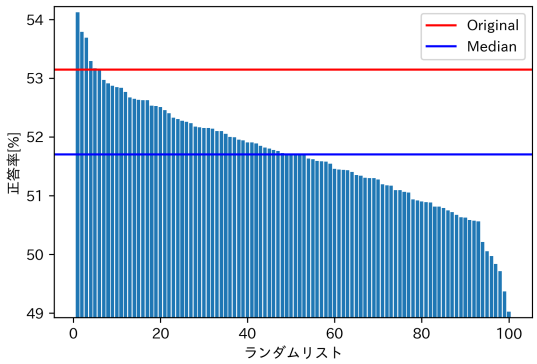

図 2: オリジナルリストによる正答率 (赤線) と ランダムリストによる正答率の中央値 (青線) の比較．重要キーワードを 3 つ以上含むニュースヘッドラインに限定した場合．ただし参照期間①の場合

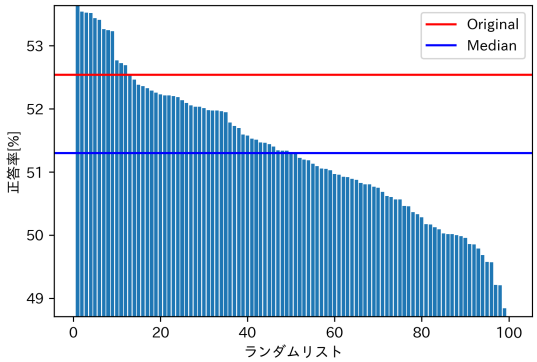

図 3: 図 2 と同様．ただし参照期間②の場合

できる．

6　アクティブ運用への活用

4 章の機械学習において 50%以上の正答率でアクティブリターン $r_c^*(d+f)$ を予測できる結果を得た．そこでアクティブ運用へ活用した場合，どの程度の収益を得られるかシミュレーションを行う．運用方法として，$r_c^*(d+f) > 0$ と予測された銘柄 c を当日 d の終値でロングし，$d+f$ 日の終値で手仕舞いする．$r_c^*(d+f) < 0$ と予測された場合は，銘柄 c を当日 d の終値でショートし，$d+f$ 日の終値で手仕舞いする．なお提案手法の有用性を確認すべく，ランダムにロングまたはショートを行う「ランダム法」とパフォーマンスを比較する．

以下の指標により，運用パフォーマンスを指標する．

累和アクティブリターン G_N

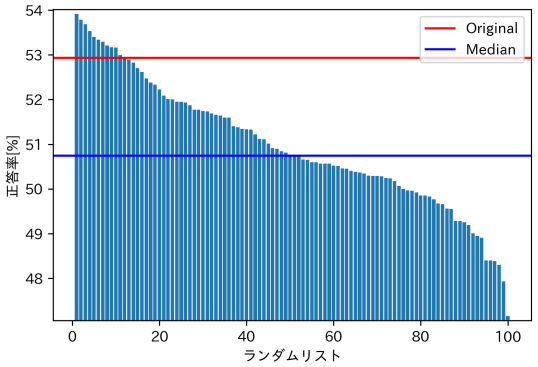

図 4: 図 2 と同様．ただし参照期間③の場合

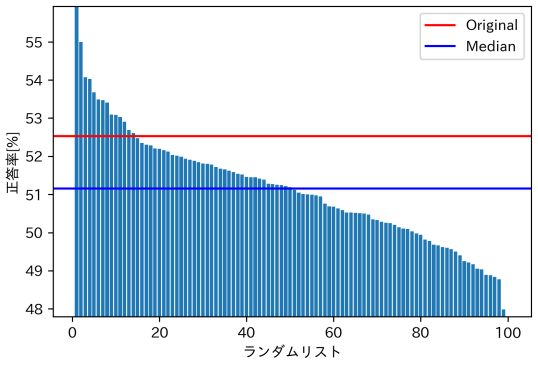

図 5: 図 2 と同様．ただし参照期間④の場合

$$G_N = \sum_{n=1}^{N} a_n r_{c_n}^*(d+f) \times 100\ [\%] \tag{7}$$

ここで，n は重要キーワードを含むニュースヘッドラインのインデックス番号であり，c_n はこのニュースに最も関連する銘柄を表す．運用行動として，$r_{c_n}^*(d+f) > 0$ と予測した場合は $a_n = 1$ (ロング) とし，$r_{c_n}^*(d+f) < 0$ と予測した場合は $a_n = -1$ (ショート) とする．なお，ニュース n に基づいて獲得したアクティブリターンを

$$G_n = a_n r_{c_n}^*(d+f) \tag{8}$$

と書く．

平均アクティブリターン $\bar{G}$

$$\bar{G} = \frac{1}{N}\sum_{n=1}^{N} G_n = \frac{G_N}{N}\ [\%] \tag{9}$$

勝率 W

$$W = \frac{|\{n|G_n > 0\}|}{N} \times 100\ [\%] \tag{10}$$

全運用回数 N に対して，プラス収益を得た運用回数の割合を示す．

インフォメーションレシオ I_R

$$I_R = \frac{\bar{G}}{\sigma_G} \tag{11}$$

ここで σ_G は G_n の標準偏差である．つまり，$\sigma_G = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(G_n - \bar{G})^2}$ である．この I_R が高いほど，リスクに対するリターンが大きい．

プロフィットファクター P_F

$$P_F = \frac{\sum\{G_n | G_n > 0\}}{\sum\{G_n | G_n < 0\}} \tag{12}$$

プラス収益の総和とマイナス収益の総和の比率であり，$P_F > 1$ であれば効率的な投資運用であるといえる．

ペイオフレシオ P_R

$$P_R = \frac{\langle\{G_n | G_n > 0\}\rangle}{\langle\{G_n | G_n < 0\}\rangle} \tag{13}$$

プラス収益の平均値とマイナス収益の平均値の比率であり，$P_R > 1$ であれば 1 取引あたりの利益が損失よりも大きいことを示す．

運用パフォーマンスを表 4 と表 5 に示す．本研究ではニュースヘッドラインに関係する全銘柄を運用対象とするため運用回数が多い．そのため累和アクティブリターン G_N は大きいものの，1 運用あたりの平均アクティブリターン $\bar{G}$ は小さい．また前述と同様に，3 つ以上の重要キーワードを含むニュースヘッドラインに限定した方が勝率 W が高く，I_R や P_F などの指標からも運用の効率性が向上していることを確認できる．

図 6〜図 9 に，累和アクティブリターン G_N の推移を示す．なお重要キーワードを 3 つ以上含むヘッドラインを対象にした場合である．ランダムにロングまたはショートを行う「ランダム法」に比べ，提案手法 (運用コスト $C = 0\%$) の優位性を確認できる．したがって本提案手法による運用判断は妥当であると言える．また，表 5 において勝率 W は 53%程度にも関わらず，運用回数が多いため順調にアクティブリターンを積み上げている様子を確認できる．しかし運用回数が多いほど，運用コストも大きくなる．そこで，式 (7) に運用コスト C を差し引くことで G_N を計算する．

$$G_N = \sum_{n=1}^{N} \left[a_n r^*_{c_n}(d+f) - C\right] \times 100 \ [\%] \tag{14}$$

結果として，運用コストが $C < 0.10\%$ (10bp 未満) であれば超過収益を維持できる可能性がある．

表 4: アクティブ運用において，重要キーワードを 1 つ以上含むヘッドラインに限定した場合

期間	G_N	$\bar{G}$	W	I_R	P_F	P_R
①	2836%	0.036%	52.79%	0.013	1.047	0.94
②	2701%	0.037%	52.64%	0.013	1.049	0.94
③	3024%	0.043%	52.81%	0.015	1.056	0.94
④	2180%	0.035%	52.58%	0.012	1.045	0.94

表 5: 表 4 と同様．ただし重要キーワードを 3 つ以上含むヘッドラインに限定した場合

期間	G_N	$\bar{G}$	W	I_R	P_F	P_R
①	1988%	0.094%	53.16%	0.030	1.113	0.98
②	1491%	0.068%	52.50%	0.023	1.085	0.98
③	1921%	0.104%	52.96%	0.036	1.133	1.01
④	1550%	0.089%	52.79%	0.029	1.108	0.99

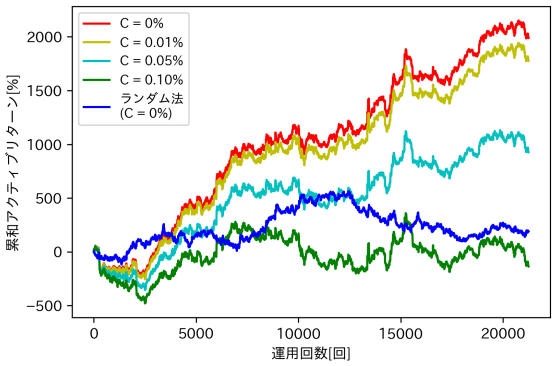

図 6: 参照期間①における累和アクティブリターンの推移．ただし重要キーワードを，3 つ以上含むニュースヘッドラインを対象とした場合

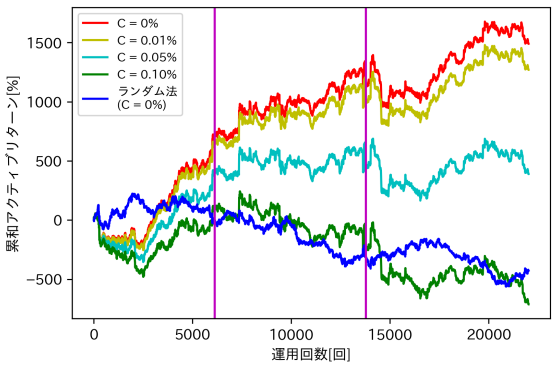

図 7: 図 6 と同様．ただし参照期間②の結果．縦線はキーワードリストの更新タイミング

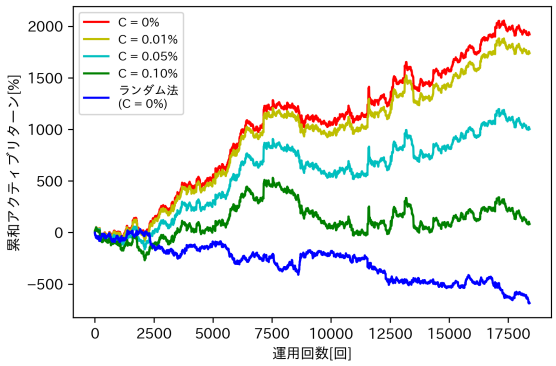

図 8: 図 6 と同様．ただし参照期間③

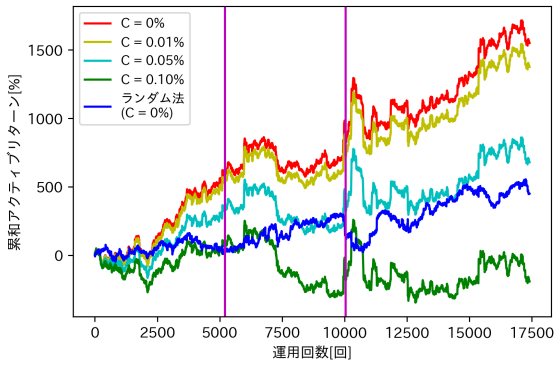

図 9: 図 6 と同様．ただし参照期間④の結果．縦線はキーワードリストの更新タイミング

7 まとめ

本研究ではニュースヘッドラインを解析対象とし，株価のボラティリティを高めるサプライズワードを抽出し，それらの組み合わせと翌日のアクティブリターンの関係を機械学習した．統計的な比較実験を通じてボラティリティーに着眼する妥当性を確認し，アクティブ運用においても 10bp 未満の運用コストならば超過収益を獲得できる可能性を得た．

本研究の一部は，文科省科研費基盤研究 (C) (No. 16K00320) の助成により行われました．

参考文献

[1] 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一: “意見抽出のための評価表現の収集,” 自然言語処理, vol.12, no.3, pp.203–222, 2005.

[2] 五島圭一, 高橋大志: “株式価格情報を用いた金融極性辞書の作成,” 自然言語処理, vol.24, no.4, pp.547–577, 2017.

[3] C. Kearney and S. Liu: “Textual Sentiment in Finance: A Survey of Methods and Models,” *International Review of Financial Analysis*, vol.33, pp.171–185, 2014.

[4] 株式会社 QUICK: https://corporate.quick.co.jp (参照日 2020.2.23)

[5] MeCab: https://taku910.github.io/mecab/ (参照日 2020.2.23)

[6] C. Cortes and V. Vapni: “Support Vector Networks,” *Machine Learning*, vol.20, no.3 pp.273–297, 1995.