アナリストレポートと企業業績の関係解析 (第一報)

Analysis of relationships between analyst reports and corporate performances (Preliminary Result)

北島良三 [1*]　酒井浩之 [1]　上村龍太郎 [2]
坂地泰紀 [3]　平松賢士 [4]　栗田昌孝 [4]
Ryozo Kitajima[1]　Hiroyuki Sakai[1]　Ryotaro Kamimura[2]
Hiroki Sakaji[3]　Kenji Hiramatsu[4]　Masataka Kurita[4]

[1] 成蹊大学
[1] Seikei University
[2] 東海大学
[2] Tokai University
[3] 東京大学
[3] The University of Tokyo
[4] 株式会社アイフィスジャパン, 株式会社金融データソリューションズ
[4] IFIS JAPAN LTD.，Financial Data Solutions, Inc.

Abstract: In this paper, we try to analyze relationships between analyst reports and corporate performances. The analyst reports are documents written about markets forecasts and they are useful for investment judgment. As analyst reports are written in natural language and data to be analyzed becomes complicated, a neural computational method called 'potential learning' which can interpret internal representations was used. As a result, we found that a generalization performance of the model was 0.6773 (accuracy) and words related to 'word category: abstraction' may affect corporate performances.

1　はじめに

本研究は，アナリストレポートと企業業績の関係について解析を行ったものである．アナリストレポートは，証券アナリストがレポート対象企業について，企業概要や事業概要，そして今後の業績予想などをまとめたレポートであり，投資判断に有用である．

このアナリストレポートであるが，記述内容・表現にアナリストの個性・心情が現れると考えられる．例えば確固たる根拠により業績上向きの判断とならない場合と，現状では上向くとは言えないものの，何らかの兆候を感じている場合には使用表現に差が生じてくることが考えられる．本研究では，このアナリストレポートに記載されている内容と企業業績の間に，例えば特定表現の出現が業績変化に影響を与えるか?といった関係性の有無について解析を試みた．

関連研究として太田らは経営者予想とアナリスト予想の精度とバイアスについて調べ,「アナリスト予想は，経営者予想の精度とバイアスに大きな影響を受けていた」[1] と報告している．また，近藤らはアナリストレポート中の株式推奨と利益予想について関連性を調査しており，例えば株式推奨変更がない場合において市場の反応の方向は「利益予想の修正方向によって決まる」[2] と報告している．近藤らの研究ではアナリストレポートから得た数値を解析に用いているが，本研究では数値データではなく文字データを解析対象としている点で異なっている．

*連絡先：成蹊大学理工学部情報科学科
〒 180-8633 東京都武蔵野市吉祥寺北町 3-3-1
E-mail: r-kitajima@st.seikei.ac.jp

2　解析手法

2.1　解析の流れ

本研究は，1. アナリストレポートから予想根拠情報 (後述する) を抽出する．2. アナリストレポートと企業業績の関係を解析するための解析用データを作成する．3. アナリストレポートを入力として，業績の増減を出

力する分類器を作成する．4. 解析結果を解釈する．という 4 つの流れにより実施される．以下各手順に沿って概要を述べる．

2.2 アナリストレポートからの予想根拠情報の抽出

はじめにアナリストレポートより「アナリスト予想根拠情報」を抽出する．これはアナリストレポートから情報抽出を行った小林らの研究により「アナリストレポートの内容を把握するうえで重要な，アナリスト予想の根拠情報を含む文」[3] と定義される文であり，アナリストレポートの核心をなす記述と言える．本研究ではこのアナリスト予想根拠情報を対象に，企業業績との関係を解析していく．アナリスト予想根拠情報は小林らの研究手法により抽出した．

抽出されたアナリスト予想根拠情報は銘柄 (証券コード) 毎にレポート発行年月順にまとめられ，本研究では 2016 年度決算期間 (例えば 2017 年 3 月が決算月の企業の場合は 2016 年 4 月から 2017 年 3 月まで) に発行されたアナリストレポートを解析対象として用いた．またこれは後述する企業業績データと組み合わせた結果，440 銘柄に対して記述されたアナリストレポートが解析対象となった．組み合わせた企業業績データであるが，これには売上高を使用した．より正確には 2016 年度売上高と 2015 年度売上高の差を求め，2016 年度売上高の方が 2015 年度売上高よりも高い銘柄にはターゲットフラグとして「1」を，そうでない銘柄に「0」をそれぞれ割り当てた．

2.3 解析用データの作成

抽出されたアナリスト予想根拠情報は文字列データであるため，形態素解析を行わないと形態素情報を得ることができない．そこでアナリスト予想根拠情報に対して形態素解析を実施し，解析用データを作成した．形態素解析には日本語形態素解析システムである JUMAN を用い,「単語カテゴリ」と「単語ドメイン」を得た．両者は JUMAN の辞書に登録されている各単語に関する分類情報であり，全部で 22 種類のカテゴリと 12 種類のドメインが登録されている．例えば単語「時計」は，カテゴリ「人工物-その他」，ドメイン「家庭・暮らし」となる．

ドメイン並びにカテゴリーは各銘柄に対するアナリストレポートでの出現頻度 (Term Frequency, TF) を，全アナリストレポートでの出現状態 (文書頻度逆数, (Inverse Document Frequency, IDF)) により重み付けした値 (TF-IDF) として記録し解析用データとした．ただし銘柄数 440 の半分よりも多い銘柄数で TF-IDF 値が 0 となっているものを除去し，表 1 に示す 26 のカテゴリとドメインを最終的に解析に用いた (変数番号 15 は本研究用に追加したカテゴリである)．

2.4 業績の増減を出力する分類器の作成

アナリスト予想根拠情報は言語データであるため解析データは複雑なものとなる．そこで，解析には複雑なデータの解析に定評のあるニューラルネットワークを使用した．しかし，ニューラルネットワークはブラックボックスと称されるほど内部表現の解釈が困難であり [4]，どの入力変数が学習に活用されたのかを理解することは容易ではない．そのため，本研究では重要変数の抽出が可能な「潜在学習」[5] と呼ばれるニューラルネットワークを解析に採用した．潜在学習はこれまでに北島らによって，太陽風を対象とした研究 [6] (数値予測問題での使用)，スーパーマーケットの POS データを対象とした研究 [7] (分類問題での使用)，等に用いられており，高い予測能力と解釈性が確認されている．

潜在学習は図 1 に示されているように自己組織化マップ (Self-Organizing Maps, SOM) と多層パーセプトロン (MultiLayer Perceptron, MLP) が基となっている二段階の学習から構成されている．図中の ① は知識獲得段階と呼ばれる段階で，入力ニューロンの潜在性 (後述する) を算出し，また，SOM に基づいて知識の獲得 (学習) を行う段階である．潜在性とは「ニューロンの多様な状況に対応できる能力」と定義されるもので,「潜在性の高いニューロン (多様な状況に対応できるニューロン) は学習で重要な役割を果たすニューロンである」と解釈する．一般的なニューラルネットワークでは学習時に活用された入力を把握することは容易ではないが，潜在学習では潜在性を確認することで，活用された入力を把握することが可能である．

知識獲得段階での学習が終了すると，続いて予測段階 (図中の ② の処理) での学習が行われる．予測段階は MLP にて，そして，入力層-隠れ層間の重みの初期値に知識獲得段階で得られた，重みと潜在性より算出した値 (重み × 潜在性) がセットされ学習が行われる．通常 MLP による学習結果は初期重みに左右されるが，潜在学習ではこの初期重み設定により，獲得された知識に基づいた学習が期待できる．以上が潜在学習の概要である．なお，潜在学習にはいくつかのバリエーションがあり，ネットワーク構造や潜在性の算出方法などに違いがある．

本研究では，知識獲得段階において出力ユニットの数を 27 個[1]，予測段階において隠れユニットの伝達関数に双曲線正接関数を用い，ユニット数を 27 個, 出力ユニットの伝達関数にソフトマックス関数を用い，ユ

[1]解析に使用したソフトウェアである MATLAB の SOM Toolbox の設定 (mapsize:small) による．

表 1: 解析用データの変数一覧

変数番号	変数名	カテゴリ・ドメイン	変数番号	変数名	カテゴリ・ドメイン
1	人	カテゴリ	16	文化・芸術	ドメイン
2	組織・団体	カテゴリ	17	スポーツ	ドメイン
3	人工物-乗り物	カテゴリ	18	健康・医学	ドメイン
4	人工物-金銭	カテゴリ	19	家庭・暮らし	ドメイン
5	人工物-その他	カテゴリ	20	料理・食事	ドメイン
6	自然物	カテゴリ	21	交通	ドメイン
7	場所-施設	カテゴリ	22	教育・学習	ドメイン
8	場所-自然	カテゴリ	23	科学・技術	ドメイン
9	場所-機能	カテゴリ	24	ビジネス	ドメイン
10	場所-その他	カテゴリ	25	メディア	ドメイン
11	抽象物	カテゴリ	26	政治	ドメイン
12	形・模様	カテゴリ			
13	数量	カテゴリ			
14	時間	カテゴリ			
15	地名:国	カテゴリ			

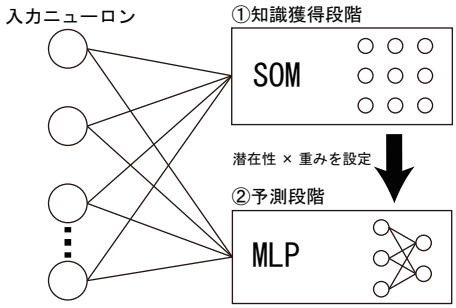

図 1: 潜在学習概要

ニット数を 2 個，とした．なお層数は，入力層 1，隠れ層 1，出力層 1 の 3 層構造である．

3 結果と考察

解析にあたり前述のデータを，学習用 (データ総数の 70%)，過学習抑制用 (データ総数の 15%)，汎化能力試験用 (データ総数の 15%)，の 3 つに分割した．さらにこれを，分割比率は変わりなく，各用途に使用されるデータサンプルがランダムに異なるものを 10 パターン作成して解析に用いた．すなわち 10 個のモデルを作成し解析を実施したわけであるが，以下述べる分類器の性能はこの 10 個のモデルより得られた値の平均値を示している．

解析の結果，アナリストレポートは正解率：0.6773, F 値：0.6244 で売上増加企業と売上減少企業に分類できることを得た ．これは潜在学習による分類と比較のために実施した，多層パーセプトロン，潜在性を使用しない潜在学習 (潜在学習とネットワーク構成は同じであるものの，学習時に潜在性を使用しないもの)，と比べ良好な結果 (結果を表 2 に示す) であったため，潜在学習がアナリストレポート解析に有用であることを確認した (潜在学習による解析の特徴として標準偏差が小さいという点も確認できた).

続いてこの分類がどの入力変数を活用して得られたものなのかについて潜在性を用いて解釈を実施する．図 2 は入力ユニットの潜在性を示したものである．この図より 11 番目のユニットの潜在性が高いことが確認できる．また潜在性と併せて入力層-隠れ層間の重みも確認したところ (図 3 (c) に潜在学習の重みを示す．この図は図形の大きさで重みの大きさを，色で重みの符号 (緑 : 正，赤 : 負) を示している)，11 番目，13 番目，14 番目，24 番目の入力ユニットに重みが集中しており，これらの入力ユニットが学習に活用されていることが確認できた．この結果は潜在性が示している入力ユニットと一致しているため，潜在性により重要入力ユニット (変数) の抽出ができていると判断した．参考までに図 3 (a) および (b) に多層パーセプトロン，潜在性を使用しない潜在学習の重みを示しているが，これらは不特定多数の入力ユニットで正負大小様々な重みを持っており，重要変数の解釈が困難 (a)，一部の入力ユニットであまり重みを持っていない等，何らかの傾向は見られるものの，傾向を絞ること (潜在学習では潜在性を確認することで入力ユニットの活用状態を把握できる) が困難 (b)，という結果であった．

なお，最も大きな潜在性を持った入力ユニットであるが，これは単語カテゴリ「抽象物」を意味するものであった．この結果より，この分類には抽象物に関する単語が重要な役割を果たしている傾向にあることが明らかになった．しかし，大きな潜在性を持つ入力ユ

表 2: 解析結果

手法	正解率		F 値	
	平均値	標準偏差	平均値	標準偏差
多層パーセプトロン	0.6727	0.0809	**0.6266**	0.0793
潜在性を用いない潜在学習	0.6545	0.0673	0.5102	0.1849
潜在学習	**0.6773**	**0.0576**	0.6244	**0.0590**

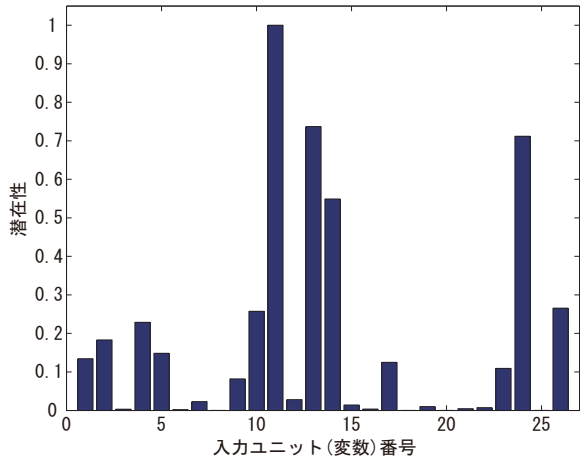

図 2: 潜在性

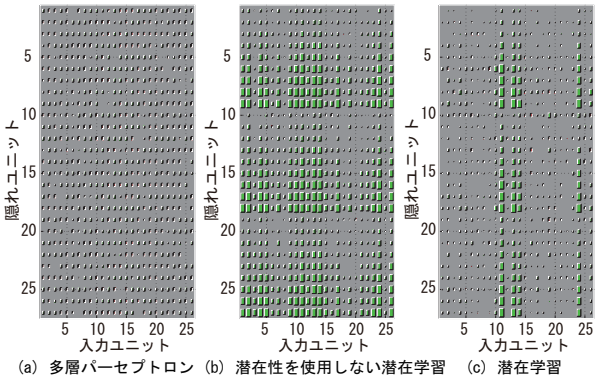

図 3: 入力層-隠れ層間の重み

ニットは 11 番目以外にも存在しているため，全体的な解釈にはこれら変数を加味する必要があり，また，具体的な単語の抽出も必要となる．これらは今後の課題である．

4 むすび

本研究は，アナリストレポートと企業業績の関係について解析を行ったものである．解析には，2016 年度決算期間に発行されたレポートから抽出されたアナリスト予想根拠情報と売上高を使用した．アナリストレポートは自然言語で記述されているため，解析データが複雑になることが予想され，解析手法として複雑なデータの解析に定評のあるニューラルネットワークのうち，内部表現を解釈可能な「潜在学習」を用いた．

解析の結果，アナリストレポートは正解率：0.6773，F 値：0.6244 で売上増加企業と売上減少企業に分類でき，この分類には「単語カテゴリ：抽象物」に関する単語が重要な役割を果たしている傾向にあることが明らかになった．

参考文献

[1] 太田浩司, 近藤江美: 経営者予想とアナリスト予想の精度とバイアス, MTEC ジャーナル, Vol. 23, pp. 33-58 (2011).

[2] 近藤江美, 太田浩司: アナリストによる株式推奨と利益予想の情報内容, 証券アナリストジャーナル, Vol. 47, No. 11, pp. 110-122 (2009).

[3] 小林和正, 酒井浩之, 坂地泰紀, 平松賢士: アナリストレポートからのアナリスト予想根拠情報の抽出と極性付与, 第 19 回金融情報学研究会予稿集, pp. 65-70 (2017).

[4] 岩崎学: データマイニングと知識発見 -統計学の視点から-, 行動計量学, Vol. 26, No. 1, pp. 46-58 (1999).

[5] Kamimura Ryotaro: Collective mutual information maximization to unify passive and positive approaches for improving interpretation and generalization, Neural Networks, Vol. 90, pp. 56-71 (2017).

[6] 北島良三, 野和田基晴, 上村龍太郎: ニューラルネットワークによる太陽風物理量を用いた地磁気擾乱指数の予測, 成蹊大学 理工学部研究報告, Vol. 54, No. 2, pp. 9-15 (2017).

[7] 北島良三, 遠藤啓太, 上村龍太郎: 入力ニューロンの潜在性に着目した小売店店舗の非継続来店顧客検知モデルの作成, オペレーションズ・リサーチ, Vol. 61, No. 2, pp. 88-96 (2016).