日本株式市場におけるテキストベース・モメンタムの実証分析

木村 友哉　　中川 慧

野村アセットマネジメント株式会社
{yy-kimura}@nomura-am.co.jp

1　はじめに

伝統的なファイナンス理論では、情報が市場に即座に織り込まれ、超過収益を獲得できないと仮定する効率的市場 [8] を礎としていた．一方で、中期的に過去の収益率が低い銘柄群 (負け組) より、過去の収益率が高い銘柄群 (勝ち組) の方が相対的に将来の収益率が高くなる「モメンタム現象」が世界中の株式市場をはじめ、あらゆる資産において観測されている [7, 1]．このようなモメンタムの発生要因として投資家の注意力に限界があることを背景とした、情報波及の遅延による株価の過少反応を挙げる研究が増加している．

とりわけ、複数の銘柄間の情報波及の遅延からもたらされる株価の先行遅行関係 (リードラグ効果) を利用した株価の予測可能性に関する多くの実証研究が存在する．例えば、産業内における情報波及遅延によるリードラグ効果を利用した [11] や [6] のインダストリー・モメンタム、サプライチェーン・ネットワークにおけるリードラグ効果を利用した [3] や [10] によるカスタマー・モメンタム、テキストデータを用いたネットワークにおけるリードラグ効果を利用した [4, 5] によるテキストベース・インダストリー・モメンタム等がある．

しかしながら、これらの研究においては [4, 5] のテキストデータを除き、投資家からの注目度の高いデータを利用している．そこで本稿では、[4, 5] に倣い、日本の株式市場に上場する各企業の事業内容を記述したテキストデータから、情報波及の遅延が発生するであろう投資家の注目度が低い類似企業群 (テキストベース業種分類) を特定し、当該企業群におけるインダストリー・モメンタム効果を検証する．その際に、[4, 5] においては、Bag of Words(BOW) をベースとした各企業の特徴ベクトルを利用しているが、本稿では BOW 手法に加えて、word2vec をベースとした Sparse Composite Document Vector[9] を利用し、企業の特徴ベクトルを算出する．

2　有価証券報告書を用いた業種分類

2.1　有価証券報告書

有価証券報告書は当該事業年度における最新の事業内容や事業環境に対する経営者の認識や包括的な企業業績を投資家に開示する法定書類[1]である．

ファイナンス分析では、主に財務諸表などの数値データが利用されるが、本稿では有価証券報告書の内、「事業の内容」、「経営方針、経営環境及び対処すべき課題等」、「事業等のリスク」、「経営者による財政状態、経営成績及びキャッシュ・フローの状況の分析」、「研究開発活動」に相当する項目のテキストデータを利用する．そしてこれら事業内容に関する記述から各企業の特徴ベクトルを作成し、クラスタリングによってテキストベースの業種を作成する．対象企業は東京証券取引所の第一部上場銘柄（TOPIX 構成銘柄）とし、2010 年から 2018 年に EDINET 上に提出された有価証券報告書を取得した．

2.2　有価証券報告書のベクトル表現

取得したテキストデータを Mecab を用いて形態素解析（分かち書き）を行い、テキストデータから単語リストに変換する．

本稿では [4, 5] で用いられた (1)Bag of Words 手法[2] に加えて、(2)Wikipedia 日本語から学習した word2vec モデルを応用した SCDV[9] を利用したベクトル化も行う．

2.3　Sparse Composite Document Vector

Wikipedia 日本語版をコーパスとして word2vec で構築した 300 次元のベクトル空間上の単語ベクトルに対して、GMM を用いてクラスタ数 30 としたソフトクラスタリングを適用する．各単語について単語

[1] 上場する有価証券発行企業は、流通市場に向けた情報開示のために、各事業年度終了後 3 か月以内に金融庁・財務局のシステムである EDINET を通じて内閣総理大臣に提出することが金融商品取引法によって義務付けられている．

[2] Bag-of-words では各単語を one-hot ベクトルに変換し, 各単語ベクトルの総和を文書ベクトルとする．

ベクトルに各クラスタへの所属確率を乗じて計算した word-cluster vector を結合することで word-topics vector (wtv_i) を計算する. これを各単語の特徴を表すベクトルとして、文書中に登場する単語の wtv_i の総和を計算し、雛形となる文書ベクトル dv を得る. 最後に dv の要素値のうち、閾値パラメタ (1%) を下回る要素について 0 に置換したベクトルを、文書ベクトル $SCDV$ として得る.

2.4 クラスタリングによる業種の作成

BOW および SCDV で作成した各銘柄の文書ベクトルに対してクラスタリングを適用することで、テキストベースの業種分類を作成する. 具体的には各ベクトルに対して k-means を用いてクラスタリングを行う. 代表的な業種分類である東証 33 業種の比較可能性の為クラスタ数を 33 とした. 本稿では作成した業種をそれぞれ BOW33, SCDV33 と略称する.

2.5 伝統的な業種分類

日本市場の投資家に利用される代表的な業種分類として、東証 33 業種と Global Industry Classification Standard(GICS) が挙げられる.

日本の代表的な証券取引所である東京証券取引所に上場する全ての株式は、証券コード協議会によって定められる 33 業種に分類されている. 東京証券取引所では同協議会の定める業種を採用し、業種別株価指数などを公表していることから、この業種分類は一般に東証 33 業種と呼ばれる.

また GICS は米国の Standard & Poors と Morgan Stanley Capital International(MSCI) の開発した国際的な業種分類である. GICS は業種の粒度に応じて 4 階層の分類が構成されており、本稿では東証 33 業種に比較的近い業種数である GICS24 産業グループを用いる.

本稿ではそれぞれ T33, GICS24 と略称する.

2.6 業種分類の分類力比較

[2] は「同業種銘柄とのリターン相関の平均値」と「異業種銘柄とのリターン相関の平均値」の格差を比較することによって、米国の代表的な業種分類のパフォーマンス比較を行った. 本稿では [2] の手法を用いてファイナンス実務および学術研究で利用される日本の代表的な業種分類およびテキストベース業種である SCDV33 および BOW33、さらにランダムに業種を割り当てたランダム業種分類の比較検証を行った.

結果は、T33 の分類能力が最も高く、以降は GICS、SCDV33、BOW33 という順 (図 1) であった.

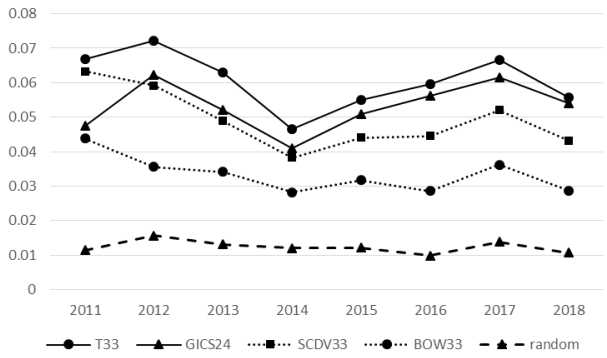

図 1: 各業種分類による分類能力の比較

表 1: 2 標本に差異があるとする仮説検定の p 値

	T33	GICS24	SCDV33	BOW33	random
T33		0.80%	0.01%	0.00%	0.00%
GICS24	0.80%		24.04%	0.03%	0.00%
SCDV33	0.01%	24.04%		0.00%	0.00%
BOW33	0.00%	0.03%	0.00%		0.00%
random	0.00%	0.00%	0.00%	0.00%	

テキストベース業種に対して、手作業で分類が行われている伝統的な業種分類の分類能力は相対的に高い傾向にあった. 一方で「ランダムに組成した業種分類に対しては統計的有意に優れていること」、「業種数が少ないため適当な比較対象とはならないが、GICS 産業グループに対しては有意な差がみられないこと (表 1)」などから、テキストベース業種分類を用いても、銘柄の株価リターン特性に応じた分類ができていることが示された. また、SCDV33 はの分類力は [4, 5] で用いられた BOW33 よりも有意に優れていた.

3 テキストベース・インダストリー・モメンタム効果の実証分析

「伝統的業種分類の T33 と GICS24」および「テキストベース業種分類の SCDV33 と BOW33」を用いて、共通項を持つ銘柄群に関する共通情報に対する各銘柄の株価反応速度の差異を背景として発生するリードラグ効果の実証分析を行う. 具体的には [5] のテキストベース・インダストリー・モメンタムの検証方法に倣い、各銘柄の将来株価収益率について、同業種ポートフォリオの過去リターンがどれほど説明力を持つかを「クロスセクション回帰」および「分位ポートフォリオ分析」を用いて検証する. データ期間は 2011/1 から 2018/11 までとし、投資対象ユニバースは各時点に

表 2: 各説明変数の回帰係数と t 値

Industry Past Return						Own-Firm Past Return					
BOW t-1 to t-12	I-BOW t-1 to t-12	SCDV33 t-1 to t-12	I-SCDV33 t-1 to t-12	T33 t-1 to t-12	GICS24 t-1 to t-12	t-1	t-2 to t-12	Log Market Capitalization	Log Book-to-Market Ratio	R^2_adj	No. of Months/ No. of Obs
-0.01								-0.18	0.05	0.026	95
(-0.23)								(-1.80)	(0.44)		157,133
	-0.02							-0.18	0.04	0.025	95
	(-0.39)							(-1.82)	(0.37)		157,133
		0.20						-0.13	0.10	0.031	95
		(2.18)						(-1.37)	(0.97)		157,133
			0.18					-0.14	0.08	0.029	95
			(2.36)					(-1.52)	(0.72)		157,133
				0.16				-0.16	0.08	0.032	95
				(1.75)				(-1.68)	(0.75)		157,133
					0.19			-0.15	0.09	0.032	95
					(2.11)			(-1.64)	(0.86)		157,133
-0.02						-0.06	0.00	-0.16	0.04	0.048	95
(-0.26)						(-0.65)	(-0.04)	(-1.52)	(0.37)		157,133
	-0.02					-0.06	-0.01	-0.15	0.03	0.048	95
	(-0.41)					(-0.66)	(-0.06)	(-1.52)	(0.28)		157,133
		0.20				-0.07	-0.04	-0.10	0.08	0.051	95
		(2.69)				(-0.89)	(-0.38)	(-1.07)	(0.82)		157,133
			0.17			-0.07	-0.03	-0.12	0.06	0.050	95
			(2.81)			(-0.77)	(-0.31)	(-1.23)	(0.58)		157,133
				0.17		-0.07	-0.03	-0.13	0.06	0.051	95
				(2.26)		(-0.88)	(-0.35)	(-1.35)	(0.62)		157,133
					0.20	-0.07	-0.04	-0.13	0.07	0.052	95
					(2.63)	(-0.81)	(-0.38)	(-1.32)	(0.71)		157,133

おける東証一部上場銘柄とする. 本分析の目的は「情報波及の遅延によってリードラグ効果が発生する」という経済的仮説の検証をテキストベース・インダストリー・モメンタムを用いて行うものである.

3.1 インダストリー・モメンタム

本分析では T33、GICS26、SCDV33、BOW33 の各業種分類について、各銘柄の同業種時価加重ポートフォリオの過去 12 か月リターンを、インダストリー・モメンタム変数として導入する.

また投資家注目度の低い関係性を背景としたリードラグ効果をとらえるため、投資家注目度の低いインアテンショナル・インダストリー・モメンタム変数を導入する. 各銘柄について、東証 33 業種で異業種かつテキストベース業種で同業種である銘柄群の過去 12 か月リターンを、I-SCDV、I-BOW として定義する.

3.2 クロスセクション回帰による検証

検証期間における各月末時点において、銘柄月次リターンを代表的な資産価格式に IndMOM を加えて重回帰分析を行うことで得られる IndMOM の回帰係数が時系列で統計的有意に 0 から乖離するかを検証する. 尚、本手法はファイナンスでは Fama-MacBeth 回帰として知られる.

[4, 5] に倣い、式 (1) および式 (2) でクロスセクション回帰を行う. $ln(B/P)$ は企業 i の簿価時価比率の自然対数、$ln(ME)$ は企業 i の時価総額の自然対数、$MOM1$ は企業 i の過去 1 か月リターン、$MOM12$ は企業 i の過去 12ヶ月リターン (過去 1ヶ月リターンを除く)、そして $IndMOM12$ を過去 12ヶ月の産業リターン、すなわちインダストリー・モメンタム変数とする. また回帰の前処理として説明変数に対して正規化および上下 5%点ウィンザー化を行う.

$$R_{i,t} = \gamma_t^0 + \gamma_t^1 ln(B/P)_{i,t} + \gamma_t^2 ln(ME)_{i,t} + \gamma_t^5 IndMOM12_{i,t} \tag{1}$$

$$R_{i,t} = \gamma_t^0 + \gamma_t^1 ln(B/P)_{i,t} + \gamma_t^2 ln(ME)_{i,t} + \gamma_t^3 MOM1_{i,t} + \gamma_t^4 MOM12_{i,t} + \gamma_t^5 IndMOM12_{i,t} \tag{2}$$

回帰係数である γ と、その回帰係数の符号の有意性を判断する指標となる t 値について、それぞれ示したのが表 2 である. 尚、t 値は系列相関に頑健な newey-west 修正後の標準誤差を用いて算出している.

結果は、まず伝統的な業種分類に関して、[11] や [6] の先行研究で示されるように、日本市場においてもインダストリー・モメンタムが観測された. 一方でテキストベースの業種分類によるインダストリー・モメンタムは、BOW による手法では有意性は確認されなかったが、SCDV による業種分類においては伝統的業種分類よりも高い有意性が観測された. また、投資家注目度が低いと考えられる I-SCDV では、さらに高いインダストリー・モメンタムが観測された.

3.3 10 分位ポートフォリオによる検証

与えられた銘柄ユニバースについて、注目する特性値に関して 10 分位点で区切った 10 の銘柄群の時価総額加重ポートフォリオを作成する. 高位ポートフォリオ－低位ポートフォリオのリターン格差を特性値リターンとし、その推移を考察する. 本稿では、東証一

部上場銘柄をユニバース、業種ポートフォリオの過去リターンであるインダストリー・モメンタムを特性値として、業種分類ごとに特性値リターンを計算する. 尚、リバランスは毎月末に行うとし、取引手数料は考慮していない.

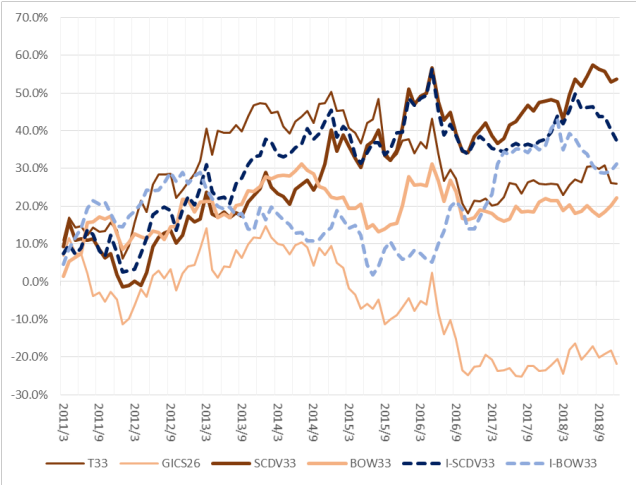

図 2: 分位ポートフォリオの格差リターン

この期間においては SCDV33 の特性値リターンが最も高く、次点に I-SCDV>I-BOW>T33>BOW33>GICS24 という結果 (図 2) となり、テキストベース業種分類が伝統的な業種分類をアウトパフォームする傾向にあった. ただし、本検証期間においてはどの業種分類においても特性値リターンの市場リスク調整後アルファ(CAPM α) は統計的有意に 0 からかい離していない. また、分位分析では各分位ポートフォリオに単一業種が集中する傾向にあり、かつ業種に他の様々なリスクが相関していることから、本結果の解釈には注意が必要である.

4 まとめと今後の課題

本稿では BoW および SCDV を用いて、有価証券報告書のテキストデータの特徴ベクトルを作成し、クラスタリングを適用することでテキストベースの業種分類を作成した. 株価リターンを用いた分類力指標では、伝統的な業種分類である東証 33 業種や GICS よりは低位であったが、銘柄の株価リターン特性に応じて有意に分類できていることを確認した. さらに各業種分類において同業種ポートフォリオの過去 12 か月リターン（テキストベース・インダストリー・モメンタム変数）と銘柄将来リターンとの関係性を検証することで、「共通情報に対する情報波及の遅延によってリードラグ効果が発生する」という経済的仮説と整合的な結果を得た.

今後の発展課題として、以下の 4 つを挙げる.

- 有報テキスト解析を用いた、テキストベース業種に対するラベリングによる解釈性の向上
- ソフトクラスタリングによる単一企業単一業種の制約撤廃、年次更新ではなく月次更新などの業種更新の高頻度化、等のより柔軟な業種分類の作成
- 任意の企業間の連結関係に注目した企業間ネットワークの作成
- 日次株価データを利用した、より短期ホライズンでの情報波及遅延およびリードラグ効果の検証

参考文献

[1] Clifford S Asness, Tobias J Moskowitz, and Lasse Heje Pedersen. Value and momentum everywhere. *The Journal of Finance*, Vol. 68, No. 3, pp. 929–985, 2013.

[2] Louis KC Chan, Josef Lakonishok, and Bhaskaran Swaminathan. Industry classifications and return comovement. *Financial Analysts Journal*, Vol. 63, No. 6, pp. 56–70, 2007.

[3] Lauren Cohen and Andrea Frazzini. Economic links and predictable returns. *The Journal of Finance*, Vol. 63, No. 4, pp. 1977–2011, 2008.

[4] Gerard Hoberg and Gordon Phillips. Text-based network industries and endogenous product differentiation. *Journal of Political Economy*, Vol. 124, No. 5, pp. 1423–1465, 2016.

[5] Gerard Hoberg and Gordon M Phillips. Text-based industry momentum. *Journal of Financial and Quantitative Analysis*, Vol. 53, No. 6, pp. 2355–2388, 2018.

[6] Kewei Hou. Industry information diffusion and the lead-lag effect in stock returns. *The review of financial studies*, Vol. 20, No. 4, pp. 1113–1138, 2007.

[7] Narasimhan Jegadeesh and Sheridan Titman. Returns to buying winners and selling losers: Implications for stock market efficiency. Vol. 48, No. 1, pp. 65–91.

[8] Burton G Malkiel and Eugene F Fama. Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, Vol. 25, No. 2, pp. 383–417, 1970.

[9] Dheeraj Mekala, Vivek Gupta, Bhargavi Paranjape, and Harish Karnick. SCDV: Sparse composite document vectors using soft clustering over distributional representations.

[10] Lior Menzly and Oguzhan Ozbas. Market segmentation and cross-predictability of returns. *The Journal of Finance*, Vol. 65, No. 4, pp. 1555–1580, 2010.

[11] Tobias J. Moskowitz and Mark Grinblatt. Do industries explain momentum? Vol. 54, No. 4, pp. 1249–1290.