中国株式市場を対象とした金融極性辞書の構築と検証

Building a Financial Polarity Dictionary Using Stock Price Information:

Analysis and Verification in Chinese Stock Markets

瞿雪吟[1]　菅愛子[1]　高橋大志[1]
Xueyin Qu[1], Aiko Suge[1], and Hiroshi Takahashi[1]
[1]慶應義塾大学大学院経営管理研究科
[1]Graduate School of Business Administration, Keio University

Abstract: This paper proposes a method of building a polarity dictionary using news articles and stock prices in the Chinese market by textual analysis in finance. In order to measure the degree of polarity, we associated the news articles' sparse composite document vectors to a score. The score is calculated by the method of event study with the abnormal change rate of stock prices on the publication date. We conducted support vector regression (SVR) and built a polarity dictionary with polarity data from learners. Furthermore, we made a comparison on accuracy to traditional ways of calculating word polarity in which news articles are represented by a one-hot wordlist. The comparison of the existed polarity is made.

1. はじめに

近年、テキストを定量化し、記事内容と株価収益率の関連性を明らかにする試みが行われている。例えば、Tetlock［1］は、文章の評価に心理学をベースとした辞書を用い、内容分析を行っている。また、Loughran/ McDonald［2］は、金融分野のテキスト分析のための辞書の提案を行っている。

中国株式市場の株式価格変動とニュースの関係を説明する多くの実証研究がなされているが、株式価格情報を用いた中国語の金融分野の単語極性辞書の研究は限定的である。また、既存の単語極性辞書で使われたテキストを定量化手法の多くは、テキスト分析の基礎的な手法である bag-of-words を用いたものであり、単語の前後の文脈が考量されてないなどの問題点が存在する。本分析では、文章のベクトル化手法として、近年、新たに提案されている Sparse composite document vector (SCDV)法（Mekala et al.[5]）を用い、テキスト分類モデルの構築、単語の極性辞書の構築を試みる。

2. 関連研究

極性辞書に関する分析は数多い。例えば、金融分野に特化した報告として、五島/高橋［3］が挙げられる。当研究では、ニュースと株式価格のデータから、キーワードリストの作成を行っており、作成した金融辞書により、将来のニュース記事や、異なるメディアのニュース記事の分類を行っている。また、関/柴本［4］は、個別銘柄の株価など対象とする金融指標によって異なる可能性があるため、金融指標の短期変動に関する語を収録した辞書を作成している。本研究では、Google 社が提供している Sparse composite document vector (SCDV)を採用し金融極性辞書の構築を試みる[1]。

3. データ

本研究においてワードリストを作成するため，ニュースは、中国金融情報サイト「和讯首页」[2]で掲載している 2013 年 1 月から 2017 年 12 月までのニュース(合計 80,325 個)を使用した。図 1 が示す通り、記事で言及された銘柄は、後ろに証券コードが付与されている。つまり、ニュースと関連する主要銘柄の情報はニュースに含まれている。

这部分个股中，涨幅最多前7只股票均为重组股或借壳股，其中多只个是因停牌早于上证综指攀升至6124高点，意味着这些股票或多或少的缺席了2007年的大牛市，重组后麻雀变凤凰，加之正常补涨，涨幅惊人，如三安光电(600703,股吧)（600703）、广弘控股(000529,股吧)（000529）、中福实业(000592,股吧)（000592）棱光实业(600629,股吧)（600629）等。

図 1　「和讯首页」掲載新聞の一例

ワードの極性評価を行うために、各銘柄の株式リターンとリスクファクター・リターンのデータを用

[1] SCDV の特徴の一つとして、テキスト分類精度が高い点が挙げられる。Dheeraj Mekala［5］は、単語の embedding にクラスタリング分析を用いることで、より複雑な文章を表現するベクトルを獲得できるとの主張を行っている。
[2] 和讯首页;http://stock.hexun.com/stocknews/index.html

いた。ファイナンシャル データオンラインサービスシステム「同花順」から、株式価格に関する日次データを取得した。また、リスクファクター・リターンのデータは、中国中央財経大学[3]が提供しているデータを用いた。

4. 作成方法

本研究における学習用データセットの作成は、大きく二つの部分に分かれる。作成過程の概略を、図 2 に示す。

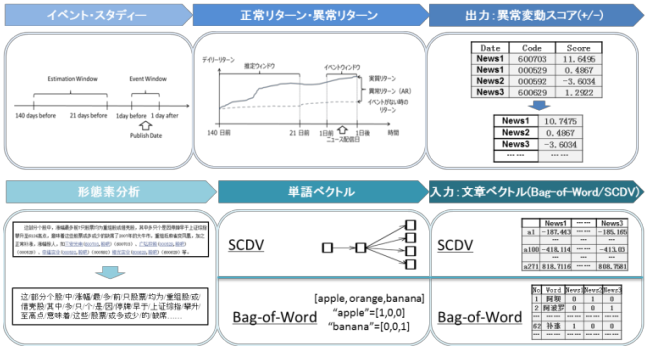

図 2　学習用データセットの作成過程

4.1 教師スコアの算出

はじめに、ニュース配信時間の調整を行った。本分析では、中国証券取引場の営業時間に合わせて、15 時以降に配信されたニュースをその翌日に調整し、また週末に配信したニュースはその次の月曜日に調整した。

次に、ニュースと関連する主要銘柄の情報を取得し、イベント・スタディにより各銘柄のリスク調整後のリターンを算出した[5]。ここで、イベント・スタディとは、企業の活動に関する情報の発表が、その企業の資産価格に与える影響を分析する手法である。株式価格変動の概念図を、図 3 に示す。

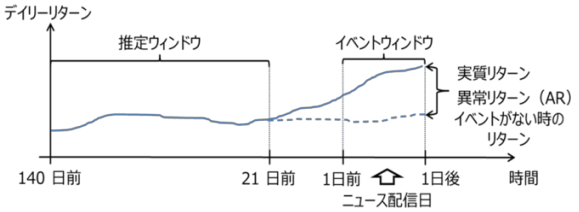

図 3　株式価格変動の概念図

[3] 中国中央財経大学;
http://sf.cufe.edu.cn/kxyj/kyjg/zgzcglyjzx/zlxzzq/98178.htm

イベント・スタディにおいては、ニュースごとに推定ウィンドウとイベントウィンドウを設定する。本分析では、推定ウィンドウはニュース配信日の 140 日前から 21 日前までの 120 日間、イベントウィンドウはニュース配信日一日前から、その翌日までとした。推定ウィンドウにおいて、Fama-French の 3 ファクターモデルにより、パラメータを推定した[6]。

4.2 ワードリストの作成

次に、2013 年 1 月から 2017 年 12 月のニュースからワードリストを作成する。

本研究では、ニュース記事内容を Bag-of-words、Sparse composite document vector (SCDV) 二つの手法によりベクトル表現を獲得し、分類モデルを構築し、それら精度の比較を行った。

4.2.1 Bag-of-Word

Bag-of-Word 手法を用いたニュースのワードリストを作成過程を、図 3 に示す。

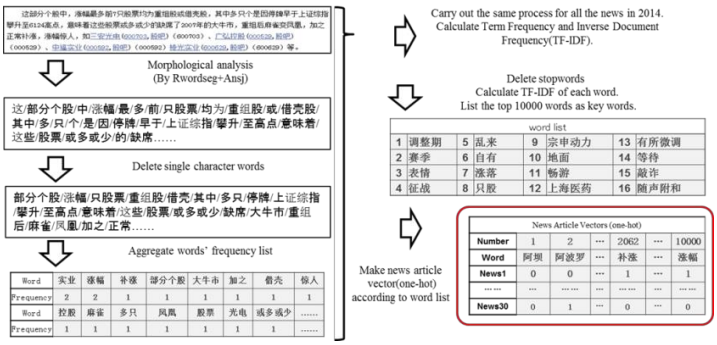

図 3 Bag-of-Word 手法を用いた
ニュースのワードリストを作成過程

本分析では、R の Rwordseg パッケージ、形態素解析ツール Ansj を用いた。金融経済分野のワードを分析するため、証券、経済、金融分野のセル辞書を導入した。セル辞書は中国で有名な中国語入力システム（IME）である Sogou[4]が提供しているものである。本分析では、1 年分のワードリスト 47,378 語の TF-IDF 値を算出し、それらのうち、上位 10,000 語を対象として分析を行った。

4.2.2 Sparse composite document vector (SCDV)

Sparse composite document vector は、グーグルが提供している文章ベクトル化手法の一つであり、高い分類精度が特徴として挙げられる（Mekala et al.[5]）。

[4] Sogou: https://pinyin.sogou.com/?r=shouji

本研究においては、Skip-gram モデルを用い、200 次元の word vector の学習を行った。GMM モデルのクラスタリング数は 60、sparsity threshold parameter は 0.33 と設定した。

4.3 分類モデル

本研究においでは、ポジティブ・ネガティブニュースの分類モデルは、SVR(Support Vector Regression)によって作成する。前節の方法で算出した教師スコアをニュース記事ベクトルに紐付け、入力をニュースの文章ベクトル、出力を、株式売買回転率により加重平均した、各関連銘柄の標準化された累積異常リターン(SCAR)として、SVR によって学習機を作成する。

5. 分析結果

分析結果について、分類モデル精度、極性、既存極性辞書との比較を記述する。

5.1 分類モデル精度

本研究は、Bag-of-words と SCDV に基づいたモデルの分類精度を比較するため、2014 年度「和讯首页」のニュース記事(サンプリング調整後 9,053 個)を対象とし、ニュースの分類モデルを作った。

本分析においては、文章の極性を基に、モデル精度の検証を行った。具体的には、文章ベクトルを学習済みの分類モデルに入力し、モデルにより推定された分類と、実際の分類が一致しているかどうかについて検証を行った。

表 1 に分析結果を示す。表の真ん中の列は SCDV を用いた分析結果であり、左の列は Bag-of-words を用いた分析結果である。モデル精度は、表 1 が示すように、SCDV を用いたニュースの分類モデルの精度が、Bag-of-words を用いたモデルより高いことを確認できる。

表 1　Bag-of-words と SCDV のモデル正解率

Sector	support vector regression models (SCDV)	support vector regression models (Bag-of-words)
In sample test (8,148 news)	72.10%	50.8%
Validation test (905 news)	70.60%	52.2%

5.2　単語の極性

SCDV 法の極性計算についで記述する。本分析では、一つの単語である word topic vector（$\overrightarrow{wtv_1}$） を、一つ文章ベクトルとして見なし、構築した分類モデルを基に単語の極性を獲得した。

5.3 既存極性辞書との比較

本節では、本研究にて構築した辞書と既存極性辞書との比較を行った。本分析では、既存辞書として Loughran McDonald Sentiment Word Lists を採用し、比較分析を行った（Loughran et al.[2]）。

比較においては、Loughran McDonald Sentiment Word Lists のポジティブおよびネガティブの単語リストを中国語に翻訳し、中国語の同義語辞書を参照し、既存辞書の単語と同じ意味を持つ単語を組み入れることで比較用の単語リストの構築を行った[5]。なお、比較分析用の Loughran McDonald Sentiment Word Lists のポジティブパートの単語数は 353 個、ネガティブパートの単語数は 374 個であった。

比較においては Bag-of-words および SCDV を採用した方法と既存辞書の比較を行った。比較においては、既存辞書に含まれる単語とそれぞれの手法によって作成された極性辞書と共通する単語、いわゆる単語の検出率（Detected Rate）の算出を行った。また、Loughran McDonald Sentiment Word Lists の単語の極性記号が、本研究で作成された極性辞書での極性記号が一致している単語の割合（Accuracy）についても算出を行った。これら二つの指標を使って、現存辞書との一致性を測定した。

表 2 は、Bag-of-Word 法により作成された極性辞書と 既存辞書との比較結果、表 3 は SCDV 法により作成された極性辞書と既存辞書との比較結果を示したものである。

表 2、表 3 より、SCDV により作成された極性辞書が、Bag-of-words により作成された極性辞書と比較して、相対的に既存辞書の共通性が高いことを確認できる。

[5] 本分析の翻訳においては Google 翻訳を用いた。

表 2　Bag-of-Word 法により作成された
極性辞書と 既存辞書との比較結果

Bag-of-Word	Loughran McDonald Sentiment Word Lists	
	Positive words	Negative words
	353	374
Detected	62	49
Positive	29	25
Negative	33	24
Detected Rate: 15.27%　Accuracy: 47.75%		

表 3　SCDV 法により作成された
極性辞書と既存辞書との比較結果

SCDV	Loughran McDonald Sentiment Word Lists	
	Positive words	Negative words
	353	374
Detected	206	216
Positive	171	173
Negative	35	43
Detected Rate: 58.05%　Accuracy: 50.71%		

6.　まとめ・今後の課題

本研究では、インターネットで掲載するされる金融市場における資産価格形成と関連性の高いニュースデータ、及び中国株式市場のマーケットデータ用い、中国語の極性辞書の構築を試みた。具体的には、Bag-of-words および Sparse composite document vector (SCDV) の手法を用い、ニュース分類モデルを構築し、単語の極性辞書の構築を行った。

分析の結果、(1)SCDV に基づいたニュースの分類モデルの精度は、Bag-of-words に基づいたモデルより高いことが確認された。(2)SCDV により作成された極性辞書は、Bag-of-words により構築された極性辞書と比較し、相対的に既存辞書と一致していることが確認された。分析手法の精緻化、分析対象データの拡張など、詳細な分析は今後の課題である。

参考文献

[1] Tetlock, P.C.: Giving Content to Investor Sentiment: The Role of Media in the Stock Market, Journal of Finance Vol.62, No.3, pp.1139-1168.(2007)

[2] Loughran, T. and B. McDonald : When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks, Journal of Finance, Vol.66, No.1, pp.35-65. (2011)

[3] 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)

[4] 関和広, 柴本昌彦: 銘柄固有の金融極性辞書の構築, 第 18 回人工知能学会, 金融情報学研究会(SIG-FIN), (2017)

[5] Mekala, D., Gupta, V., Paranjape, B., Karnick, H. SCDV : Sparse Composite Document Vectors using soft clustering over distributional representations[C], 2017 Conference on Empirical Methods in Natural Language Processing, pp.670-680,(2017).

[6] Campbell, J.Y., Lo., A.W., and MacKinlay, A.C.: The econometrics of financial markets, Princeton, NJ: princeton University press, 1997.

[7] Fama, E. F. and French, K. R.: Common Risk Factors in Returns on Stock and Bonds. Journal of Financial Economics, Vol.33 No.1, pp. 3-56. (1993).