F-047

データ分類手法を用いたブログ注目情報と株価変動の相関分析

Correlation Analysis between Blog Hot Topics and Stock Price Changes by Data Classification Method

原　慎司[1]　灘本裕紀[2]　堀内　匡[1]
Shinji Hara　Hironori Nadamoto　Tadashi Horiuchi

1. はじめに

近年，Web 上で一般の人々が容易に情報を発信する手段としてブログが注目されている．ブログは即時性・リアルタイム性のある新鮮な情報を配信しているため，新たな情報源としても注目されている．このブログを大量に収集し，ブログの集合を対象としてさまざまな手法で分析することで，一般の人々の「生の声」を抽出しようという試みであるブログマイニングと呼ばれる新しい研究が始まっている [1].

本研究では，この新しいブログマイニングに注目し，実世界の動向との相関分析の一つとして，既存のブログサーチエンジン kizasi.jp[2] を用いた株価変動と相関が高いキーワード群を抽出する枠組みについて検討する．また，得られたキーワード群を分類属性とし，データ分類手法である決定木学習とナイーブベイズ法を用いて株価変動情報の分類実験を行うことにより，抽出されたキーワード群の評価を行う．

2. データ分類手法

本研究では，データ分類手法として，決定木学習とナイーブベイズ法を用いる．以下では，それらの手法について説明する．

2.1 決定木学習

ある株銘柄 A をクラス c_1「株価上昇」とクラス c_2「株価下降」に分類することを考える．選択されたキーワード群を分類属性 $\boldsymbol{s} = (s_1, \cdots, s_n)$ として n 次元のキーワードベクトルを考え，銘柄 A が含んでいるキーワードには値 1 を，銘柄 A が含んでいないキーワードには値 0 を付けたベクトルを考え，属性ベクトルとする．これらの属性ベクトルを持つデータから構築された決定木をたどることで，クラス c_1「株価上昇」とクラス c_2「株価下降」のどちらのクラスに分類するか決定することができる．

2.2 ナイーブベイズ法

ある株銘柄 A をクラス c_1「株価上昇」とクラス c_2「株価下降」に分類する場合を考える．この銘柄に含まれるキーワード群を $\boldsymbol{s} = (s_1, \cdots, s_n)$ とすると，事後確率 $P(c|\boldsymbol{s})$ を最大とするクラスを求めることで，分類の誤りを最小とすることができる．この事後確率 $P(c|\boldsymbol{s})$ は，ベイズの定理により求めることができる．

3. 提案手法

本研究では，kizasi.jp より抽出した株銘柄注目情報を利用して株価変動に関連したキーワードを抽出する．さらに，データ分類手法である決定木学習とナイーブベイズ法によって，抽出したキーワードを用いて収集された株銘柄注目情報が適切なクラスへ分類されるか実験を行う．提案する手法の枠組みを図 1 に示す．

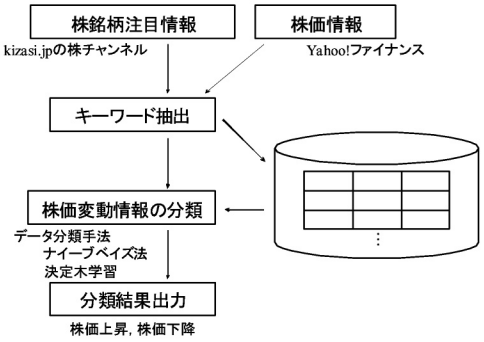

図 1: 提案手法の枠組み

表 1: 株銘柄注目情報の例

銘柄名	クラス	キーワード
I	c_1:株価上昇	上昇, 株価, 買い, スイング, ···
J	c_2:その他	下方修正, 買い, 空売り, 暴落, ···

3.1 株価変動キーワードの抽出

kizasi.jp の株チャンネルを利用して，注目されている株銘柄とそれらに関連性の高いキーワードを多数集め，それを株銘柄注目情報とする．そして，実際の株価の変動を Yahoo!ファイナンス [3] より取得し，その株銘柄注目情報を前日の株価 ($p0$) と翌日の株価 ($p1$) を用いた次式 (1)〜(3) によって，クラス「上昇」とクラス「その他」のように手動でクラス分けを行う．クラス分けされた株銘柄注目情報は，表 1 のようにデータベースに格納する．また，価格にはその日の 5 日平均値を用いる．

$$p' = \frac{p1 - p0}{p0} \quad (1)$$

$$p' \geq 0.02 \Rightarrow \text{「上昇」} \quad (2)$$

$$p' < 0.02 \Rightarrow \text{「その他」} \quad (3)$$

次に，このクラスが付与された株銘柄注目情報から，各クラスのキーワードの出現回数を求める．大規模なデータにおいて，株銘柄注目情報には，キーワードの出現頻度に大きな偏りが生じる．そこで，株価変動と相関の高いキーワードを抽出するために，閾値での打ち切りと情報エントロピーを用いた重要語の選択 [4] を行う．まず，閾値での打ち切りについては，各キーワードの出現率に基づく閾値を設け，閾値以下であるキーワードを除

[1]松江工業高等専門学校，Matsue College of Technology
[2]京都大学大学院，Kyoto University

表 2: データの概要

データ	収集期間	総銘柄情報数
A	08. 9.1〜10.31	3815（上昇: 880, その他:2935）
B	08.11.1〜12.31	5175（上昇:1519, その他:3656）
C	09. 1.1〜 2.28	6038（上昇: 803, その他:5235）

表 3: 選択されたキーワードの一部（データ A:閾値 1%）

keyword	上昇	その他	E(w)
株	392	876	0.585
東証 1 部	186	199	0.586
東 1	140	163	0.596
不動産業	84	48	0.598
JASDAQ	83	57	0.600

表 4: データセットの概要

データセット	学習データ	テストデータ
1	データ A	データ B
2	データ B	データ C

去し，キーワードを絞ることができる．予備実験として閾値毎の分類精度を比較した結果，閾値 1%において最も精度が良かった．次に，各キーワードについてのクラス毎の出現回数より，情報エントロピーの値に基づいた確率的コンプレキシティ$E(w)$ を算出し，この値が小さいものから分類属性として選択した．今回の場合，$E(w)$ が小さいキーワードw は，クラス「上昇」にクラス分けされた株銘柄注目情報によく現れる，あるいはほとんど現れない単語であることを意味する．

4. 実験

提案手法の有効性を検証するために，実際に取得した大規模データを用いて，検証実験を行った．実験のために収集したデータの概要を表 2 に示す．提案した手法によって得られるキーワードの分類能力の評価のために，学習データとテストデータに分け，分類実験を行った．選択したキーワードの一例を表 3 に示す．分類実験において，選択したキーワードは 30 個であり，分類手法としては決定木学習とナイーブベイズ法を用いた．

4.1 テストデータによる分類能力の評価

今回，収集した全データを 3 つの期間に分け，それぞれのデータを学習データ，テストデータに分けて，分類実験を行った．表 4 に示すように，先月のデータによって学習し，その翌月のデータで評価するように 2 種類のデータセットにわけた．この実験を通して，提案手法によって得られたキーワード群が新たなデータをどれほど正確に分類できるかという汎化能力を検証する．

4.2 実験結果および考察

図 2, 図 3 に決定木学習とナイーブベイズ法を用いた分類結果をそれぞれ示す．これらの図より，学習データに対するクローズド評価とテストデータに対するオープン評価を比較すると，オープン評価の方が当然低いが，精度の低下は 7%以下であり，学習データによって構築された分類モデルは，汎化能力がある程度高いものであると考えることができる．

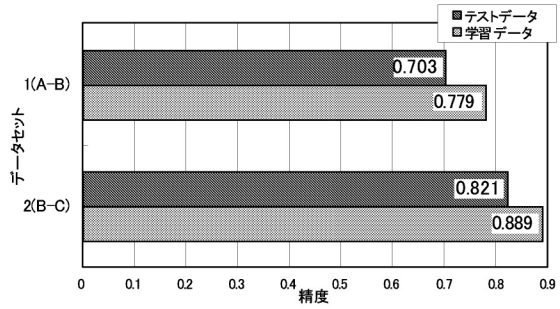

図 2: 精度の比較（決定木学習）

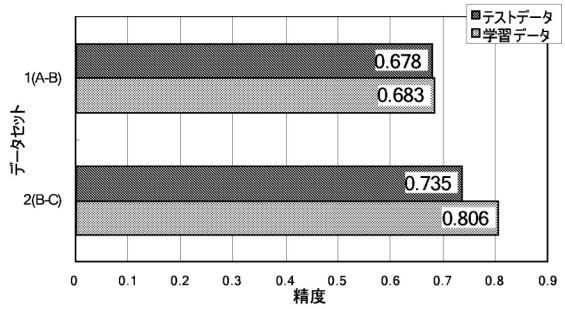

図 3: 精度の比較（ナイーブベイズ法）

また，分類精度がナイーブベイズ法よりも決定木学習の方が良かったことも結果からわかる．この理由としては，決定木学習では属性キーワード間の共起関係（依存関係）を扱うことができ，ナイーブベイズ法に比べて，分類モデルとして表現能力が高いという特徴からくるものと考えている．

5. まとめ

本研究では，kizasi.jp の株チャンネルと yahoo!ファイナンスを利用して，株価変動と相関の高いキーワード群を抽出する枠組みについて検討した．実験により，提案手法によって得られたキーワード群が，テストデータに対しても，ある程度高い分類能力を有していることがわかった．

今後は，株価以外の対象データに対しても，提案したキーワード選択法によって，意味のあるキーワードが得られるのか検証を進めていく予定である．

参考文献

[1] 奥村学，“blog マイニング－インターネット上のトレンド，意見分析を目指して－”，人工知能学会誌，Vol.21，No.4，pp.424-429，2007.

[2] 株式会社きざしカンパニー，“kizasi.jp,” ⟨http://kizasi.jp/⟩

[3] ヤフー株式会社，“Yahoo!ファイナンス”，⟨http://quote.yahoo.co.jp/⟩

[4] 竹村彰道，“統計科学のフロンティア 10 言語と心理の統計”，pp.59-128，岩波書店，2003.