NMF を用いた為替ドル円レートの変動要因分析と Adaboost を用いた予測システム

瀬川 雄基　　David Ramamonjisoa
岩手県立大学ソフトウェア情報学科ソフトウェア情報学部
g031l084@s.iwate-pu.ac.jp, david@iwate-pu.ac.jp

１．はじめに

経済の動向予測の手法として、テクニカル分析やファンダメンタルズ分析等が行われていた．しかし，数値データのみでの予測は大変困難である．よって近年ではニュース記事やソーシャルデータ等の情報と，過去の数値データを利用したテキストマイニングが行われるようになった．

テキストマイニングを行うにあたり，重要なのがテキストの選択である．日本銀行は多額の金額を市場介入することで為替の動向を調整してきた．例に挙げると，2003 年 5 月 8 日～2004 年 3 月 16 日の期間にデフレーションの克服, 円高の是正を目的に総額32 兆8694 円もの大金を介入した．また，2013 年 4 月に質的・量的金融緩和を行い 2014 年 1 月には日経平均株価が約 80%上昇, 円相場は 1 ドル 80 円から 100 円と円安となった．このような実例から，日本銀行の動向は日本の経済に大きな影響を与えることが分かる．日本銀行が定期的に公開している金融経済月報というレポートがある．そのレポートには日本銀行の会議内容も内包されており，政策の方針を知ることができる．実際にトレーダーも参考にする場合もあるため，日本銀行の動向を調べるのに有効なテキストであると考える．

本研究ではこの日本銀行の金融経済月報と，同じく定期的に公開され，日本経済について分析している日本総研の日本経済展望も併せて，2 つのレポートをテキストとし為替の動向予測を行う．

２．関連研究

参考文献 [1]では金融経済月報を用いた日経平均株価の長期予測を行っている．過去のレポートから単語を抽出し，TFIDF を求め単語ベクトルを作成．その後，SVM を用いた株価の動向を長期予測している．

参考文献 [2]では金融経済月報を用いた金融市場の分析を行っている．共起関係に基づいた単語の抽出を行い，主要単語をノードとするネットワークを構築し，主成分分析を用いて単語をグループ化する．その後，重回帰分析を行ない株価の動向を予測している．

本研究では PCA と NMF の 2 つの手法を用いて特徴抽出を行い，予測には Adaboost を用いる．

３．手法

以下では，使用する手法について解説する．

３．１　N-gram

N-gram モデルとは，ある文字列の中で N 個の文字列または単語の組み合わせを作成し，それぞれの組み合わせがどの程度出現するかを調べるモデルである．隣り合った文字列または単語の組み合わせを共起関係と呼び，その共起関係の頻度を集計した結果を共起頻度と呼ぶ．

本研究では，はじめに形態素解析を行ない，テキストを名詞のみの単語に分解している．しかし，「マイナス金利」等組み合わせにより意味を持つ文字列が分解され意味を持たない単語になってしまっている．そこで N-gram を用いて意味のある文字列を作成している．

３．２　PCA

主成分分析(Principal Component Analysis, PCA)とは，データを分析しやすいよう少数の合成変数を作成し，次元を削減するという手法である．次元を下げる理由として，機械学習や統計において，次元の呪いというデータの次元が大きすぎると認識の精度が悪くなるという現象を回避するためである．また，データを低次元に変換することにより可視化が可能になるという利点もある．参考文献 [3]では主成分分析を以下のように定義している．

“P個の変数$\{x_p\}(p = 1,2, \dots, P)$の持つ情報を，情報の損失を最小限に抑えながら，$\{x_p\}$の一次結合として与えられる互いに独立な$M(M \leq P)$個の主成分(総合的指標)$\{z_m\}$

$$z_m = \sum_{p=1}^{P} \omega_{pm}\, x_p \quad (m = 1,2, \dots, M)$$

を用いて表現する手法である．”なお，z_mは第m主成分と呼ばれ，$\omega_{pm}(p = 1,2, \dots, P; m = 1,2, \dots, M)$を結合係数とする．

３．３　NMF

NMF の説明に関しては，参考文献 [4]を参考にした．

NMF(Nonnegative Matrix Factorization, 非負値行列因子分解)とは，非負値(0 か正の値を持つ)行列を 2 つの非負値行列(特徴の行列 H，重みの行列 W)の積で近似

する手法である.

特徴の行列では，行はそれぞれの特徴であり，列がそれぞれの単語を表している．値は単語の特徴に対しての重要度を示している．本研究ではこの特徴行列を抽出し，特徴トピックとして利用する.

参考文献[5]を参考にもとの行列A(i, j)の，NMFの数式を以下に示す.

$$\mathrm{k} \in \mathrm{R}, \quad \mathrm{A} \approx \mathrm{WH}, \quad \mathrm{A} \in R^{i \times j}$$
$$\mathrm{W} \in R^{i \times k}, \quad H \in R^{k \times j}$$

NMFではこのAとWHの損失を定義し最小化する．損失関数にはユークリッド距離(EUC)を用いる.

$$D_{EUC}(A, WH) = (A - WH)^2$$

このアルゴリズムを乗法的更新アルゴリズムとして使用する．変形後の更新式を以下に示す.

$$H_{n+1} \leftarrow H_n \frac{W_n^T A_n}{W_n^T W_n H_n}, \qquad W_{n+1} \leftarrow W_n \frac{A_n H_{n+1}^T}{W_n H_{n+1} H_{n+1}^T}$$

３．３ Adaboost

Adaboost(Adaptive Boosting)とは，アンサンブル学習であるブースティングの中で最も一般的な手法であり，逐次弱学習器による学習結果から重みを調整することで，性能を向上させる手法である.

４．実験

４．１ NMFを用いた特徴トピック抽出

まず，日本総研の日本経済展望のみをコーパスとして実験を行った.

i 日本総研のホームページ[6]から日本経済展望のPDFファイルをダウンロード. PDFminerを用いてtxtファイル化.

ii Janomeを用いてファイルごとに形態素解析を行い名詞のみを抽出し，リスト化する.

iii リストの単語群から1-gram，2-gram，3-gramを作成.

iv 上記で，作成された文字列群からPythonの機械学習ライブラリである scikit-learn を用いてNMFを行い，特徴トピックを抽出.
抽出結果の一部を図1に示す.

４．２ Adaboostを用いた長期為替予測

今回は，テキストに日本総研の日本経済展望の2009年1月から2016年11月のテキストを用いた．そのテキストのうち75%を学習データに，残りの25%をテストデータとした．時系列データとして，同じく過去6年分の月次為替ドル円の上昇を1，下降を0としたデータを用いる．実装方法は参考文献[7]を参考にする.

比較するために，Adaboost と同時に Decision tree(決定株分類器)も用いた．また学習データ，テストデータそれぞれを分類し精度を確かめた．実行結果を表1に示す.

```
Topics in NMF model:
Topic #0:
基日本総研 基日本 見通し金利動向 消費税率 見通し金利 引き上げ 税率 景気分析 消費税率引き上げ 税率引き上げ 円安 金利動向 分析見通し金利 期待 物価上昇 現状景気 現状景気分析 景気分析見通し 分析見通し 労働 対策 ４月 足許 増税 インフレ 現地 程度 消費増税 給与 反動 不足 駆け込み 反動減 実施 経済対策 外需 コア 想定 緩やか ３月 数量 ６月 作用 利益 ９月 ベース 基調 経常利益 者数 展望年月
Topic #1:
もと日本 もと日本総研 展望年月 長期化 ため 調整 傾向 デフレ 計画 低迷 金融緩和 円高 悪化 受注 機械受注 押し上げ効果 統計 自動車 程度 下落 世帯 年期資料 ベース 期資料 これ 売上 維持 伸び 追加 数量 対策 期待 財政 需給 推計 ３月 ユーロ 資本 連続 労働 給与 輸出数量 ６月 ９月 マインド 輸送機械 者数 基調 持ち直し 予算
Topic #2:
復興 震災 減速 自動車 円高 億円 ユーロ ３月 予算 ため 基日本 基日本総研 統計 展望年月 計画 想定 金融緩和 場合 追加 ９月 足許 世帯 財政 機械受注 受注 企業収益 マインド 引き上げ 現地 悪化 生産指数 レート 緩やか 建設 駆け込み 利益 低迷 事業 輸送機械 規模 デフレ 期資料 金利動向 長期化 外需 反動 作用 年期資料 程度 基調
```

図１　日本経済展望のNMFによる特徴抽出結果の一部

表１：Adaboostと Decision treeの分類結果

精度	学習データ	テストデータ
Adaboost	100%	54%
Decision tree	66%	63%

５．おわりに

本研究では，日本銀行の金融経済月報，日本総研の日本経済展望の２つのテキストを用いてドル円為替の変動についてNMFとPCAによりトピック抽出を行った．図１のNMFによるトピック抽出結果から，Topic0では「金利」に関するキーワードが多くみられる．Topic1では「金融緩和」，「輸出」や「輸送」など外交に関係するキーワードがみられる．Topic2では「金融緩和」，「復興」や「震災」というキーワードがみられた.

また，Adaboostと Decision treeによる分類の結果は表１の通り，Adaboostに関して学習データへの分類は100%，テストデータの分類は54%と低かった．Decision treeに関して学習データへの分類は66%，テストデータの分類は63%となり，テストデータの分類精度はDecision treeの方が高い結果となった．この結果の要因としてはAdaboostの過学習が考えられる.

今後の課題は，Adaboost での精度を高めるために特徴抽出の調整，工夫が必要だと考えられる.

参考文献

1) 森谷英一郎：金融経済月報を用いた長期株価動向予測，岩手県立大学2015年度卒業論文(2016)

2) 和泉潔，後藤卓，松井藤五郎：テキストマイニングによる金融市場の月次動向分析，社団法人 情報処理学会研究報告(2009)

3) 加納学：主成分分析 京都大学大学院工学研究科化学工学専攻 プロセスシステム工学研究室(1997)

4) Toby Segaran：集合知プログラミング，pp253~270
5) 伊藤寛祥，天笠俊之，北川博之：論文データベースにおけるトピックの変遷の検出，DEIM Forum(2015)
6) 日本経済展望｜経済・政策レポート｜日本総研 https://www.jri.co.jp/report/medium/japan/ (2017/01/12 参照)
7) Sebastian Raschka：Python 機械学習プログラミング 達人データサイエンティストによる理論と実践 pp214~222